Multilingual Spoken Language Understanding (SLU)

601.764

2/16/2023

SLU is often 2-steps

- ASR
- NLU

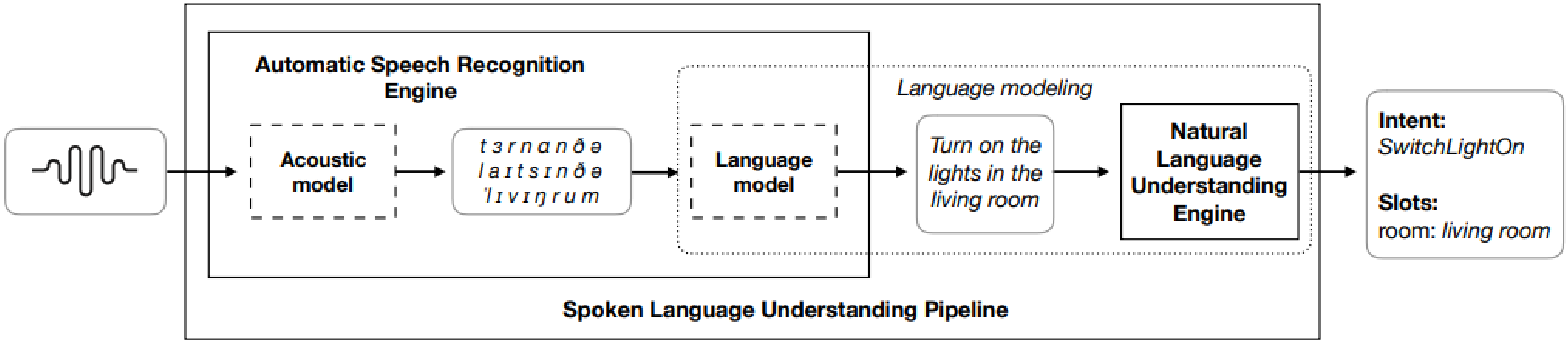


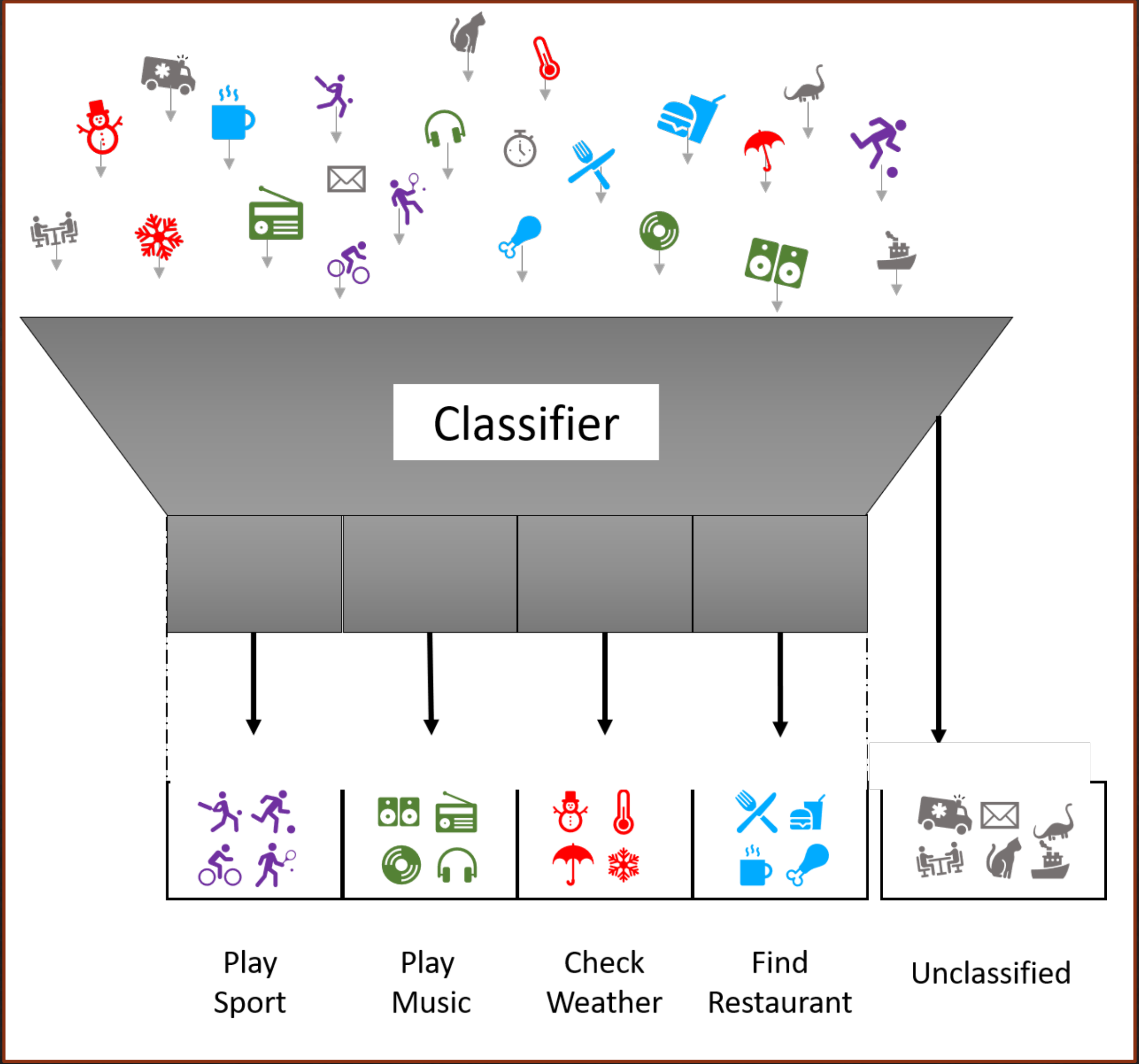
Figure 2: Spoken Language Understanding pipeline

Coucke et al., 2018

NLU (for SLU) is 2 tasks

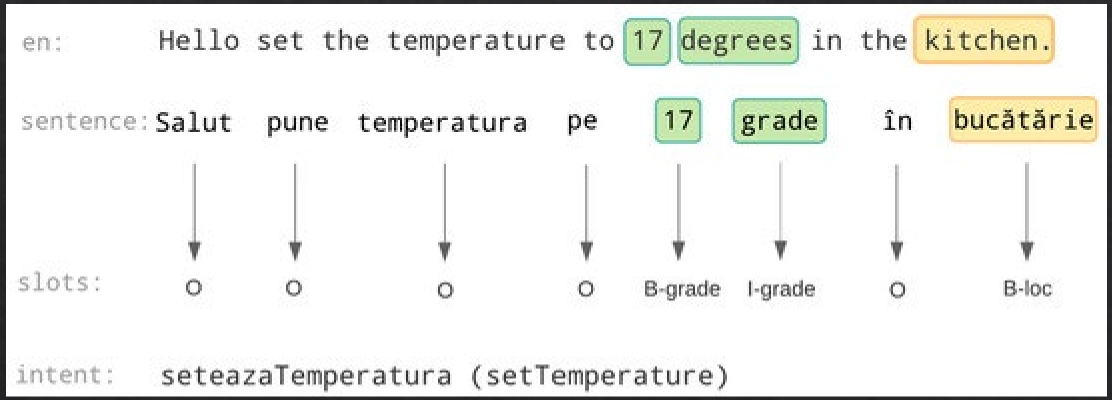
- Intent Detection
- Slot Filling

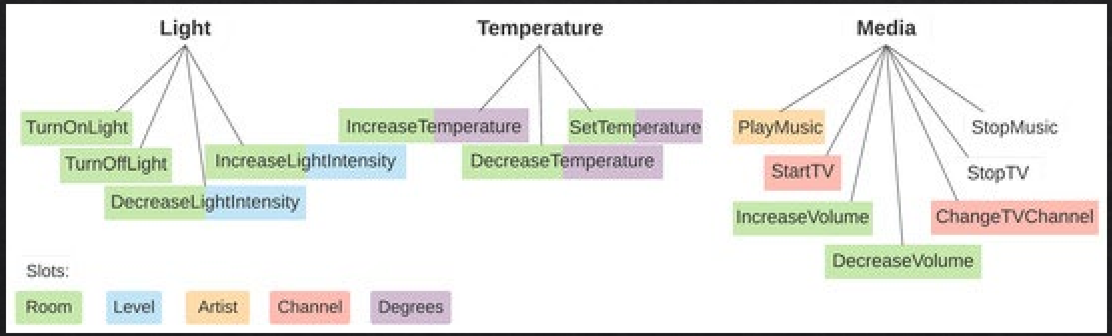
Intent Detection



https://docs.smartly.ai/docs/intent-detection

Slot Filling





Stoica et al., 2021

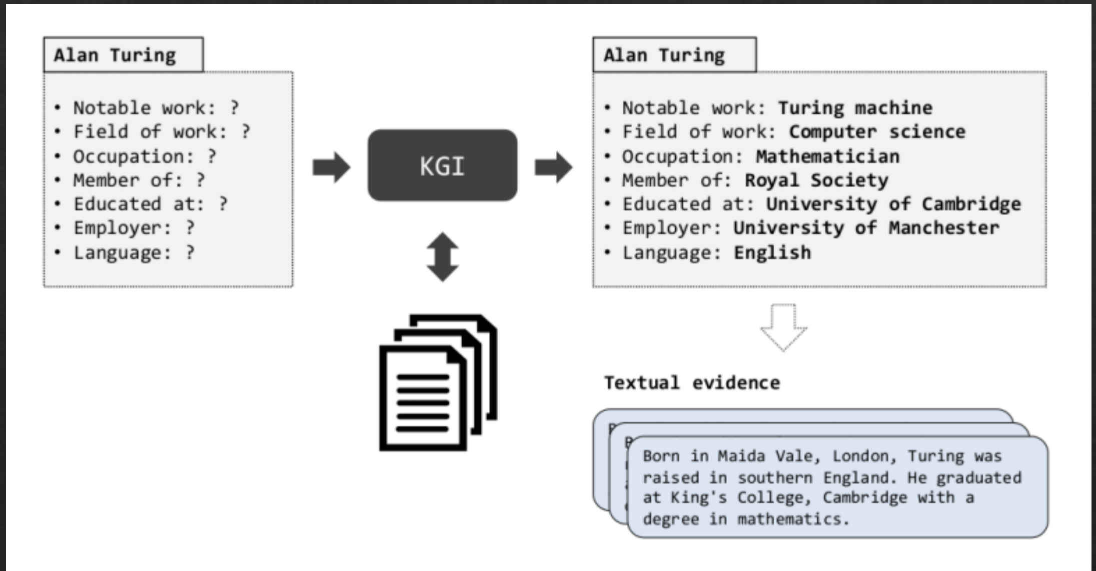


Figure 1: Slot Filling task

Glass et al., 2021

https://paperswithcode.com/task/slot-filling

Datasets (2 things again)

- ATIS
- SNIPS

How may I help you?

- Call types
- Routing
- Not trying to explicitly determine arguments
- Orthographically Transcribed
- 10,000 transactions

N·H
ELSEVIER

Speech Communication 23 (1997) 113–127

SPEECH
COMMUNICATION

How may I help you?

A.L. Gorin [*], G. Riccardi [1], J.H. Wright [2]

AT&T Labs-Research, Florham Park, NJ, USA

Received 14 February 1997; revised 23 May 1997

SNIPS

- 2018
- English, French, German
- Spanish, Korean (limited)
- More added continuously

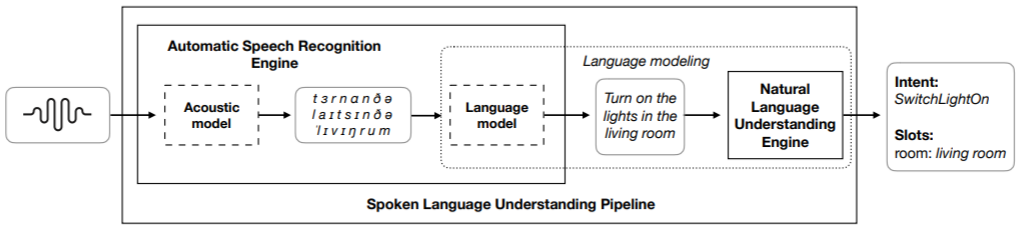


Figure 2: Spoken Language Understanding pipeline

Snips Voice Platform: an embedded Spoken Language Understanding system for private-by-design voice interfaces

Alice Coucke **Alaa Saade** **Adrien Ball**
Théodore Bluche **Alexandre Caulier** **David Leroy**
Clément Doumouro **Thibault Gisselbrecht** **Francesco Caltagirone**
Thibaut Lavril **Maël Primet** **Joseph Dureau**

Snips
Paris, France

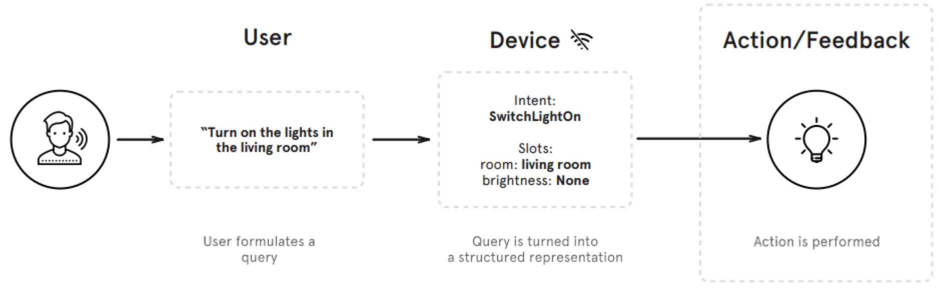


Figure 1: Interaction flow

Airline Travel Information System

- DARPA, early 90's

Utterance	***How much is the cheapest flight from Boston to New York tomorrow morning?***
Goal:	Airfare
Cost_Relative	*cheapest*
Depart_City	*Boston*
Arrival_City	*New York*
Depart_Date. Relative	*tomorrow*
Depart_Time. Period	*morning*

Is ATIS still useful?

- Error Analysis
- Says error rates of less than 5% call for other datasets such as:
  - French Media Dialogue 3x size, 10% error (though WoZ as a downside)
  - Let’s Go (Pgh Bus, though SLU not yet available …. 2005)

WHAT IS LEFT TO BE UNDERSTOOD IN ATIS?

Gokhan Tur   Dilek Hakkani-Tür   Larry Heck

Speech at Microsoft | Microsoft Research
Mountain View, CA, 94041
gokhan.tur@ieee.org dilek@ieee.org larry.heck@microsoft.com

2010

Semantic annotation of the French Media dialog corpus ◇

H. Bonneau-Maynard[1], S. Rosset[1], C. Ayache[2], A. Kuhn[2], D. Mostefa[2] and the MEDIA *consortium*

(1) LIMSI-CNRS/FRANCE, (2) ELDA/FRANCE

{maynard,rosset}@limsi.fr, {ayache,kuhn,mostefa}@elda.org, media@elda.org

r	word seq.	mode	attribute name	attribute value
0	euh	+	null	
1	oui	+	response	oui
2	l’	+	refLink-coRef	singulier
3	hôtel	+	BDObject	hotel
4	dont	+	null	
5	le prix	+	object	paiement-montant-chambre
6	ne dépasse pas	+	comparative-payment	inferieur
7	cent dix	+	payment-amount-integer-room	110
8	euros	+	payment-unit	euro

Figure 1: Example of the semantic attribute/value representation for the sentence *“uhm yes the hotel whose price doesn’t exceed one hundred and ten euros”*. The relations between attributes are given by their order in the representation and the composed attribute names. The segments are aligned on the sentences.

	wizard	client
#utterances	19633	18801
mean #words per utterance	14.4	8.3
vocabulary size	1932	2715

Table 1: Main characteristics of the 1257 dialog MEDIA corpus. The average dialog duration is 3’30.

Multi-ATIS

- English ATIS → {Turkish, Hindi}
- 2018
- Name?

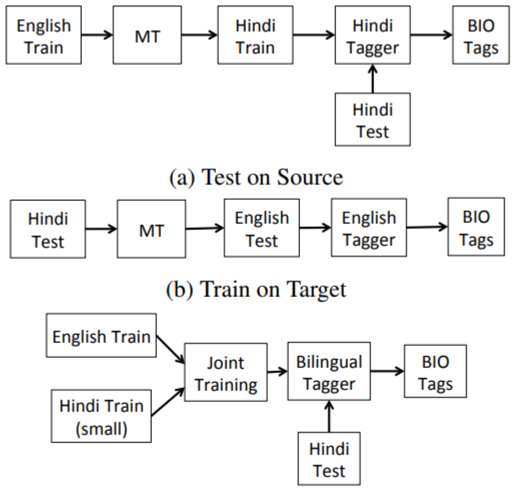
Utt: find a one way flight from boston to atlanta on wednesday
Slots: O O B-RT I-RT O O B-FC O B-TC O B-DDN

(a) English Utterance

Utt: बुधवार को बोसटन से अटलांटा तक जाने वाली एकतरफ़ा उड़ाने खोजें
Slots: B-DDN O B-FC O B-TC O O O B-RT O O

(b) Hindi Utterance

Fig. 1: English and corresponding Hindi utterance and their slots in BIO encoding. The correct intent label is "flight". Tags: RT - *round trip*, FC - *from city*, TC - *to city*, DDN - *departure day name*.



(c) Our Approach

Fig. 2: Former approaches for Cross-lingual SLU compared to our approach of joint training across languages. Note that our approach does not rely on a Machine Translation (MT) step, which may be unreliable for relatively low resource languages.

(ALMOST) ZERO-SHOT CROSS-LINGUAL SPOKEN LANGUAGE UNDERSTANDING

*Shyam Upadhyay**[1] *Manaal Faruqui*[2] *Gokhan Tür*[2] *Dilek Hakkani-Tür*[2] *Larry Heck*†[3]

[1] University of Pennsylvania, Philadelphia, PA
[2] Google Research, Mountain View, CA
[3] Samsung Research, Mountain View, CA
shyamupa@seas.upenn.edu, {mfaruqui, gokhant, dilekh}@google.com, larry.heck@ieee.org

Multi-ATIS++

- Professional Translators from English
- Spanish, German, French, Portuguese, Chinese, Japanese
- 2020

End-to-End Slot Alignment and Recognition for Cross-Lingual NLU

Weijia Xu* University of Maryland weijia@cs.umd.edu

Batool Haider Amazon AI bhaider@amazon.com

Saab Mansour Amazon AI saabm@amazon.com

EN	show		departures	from	atlanta	for	american
	O		O	O	B-fromloc.city_name	O	B-airline_name
ES	Muestra		salidas	desde	Atlanta	de	American
	O		O	O	B-fromloc.city_name	O	B-airline_name
PT	Mostre		partidas	de	Atlanta	da	American
	O		O	O	B-fromloc.city_name	O	B-airline_name
DE	Zeige		Abflüge	von	Atlanta	für	American
	O		O	O	B-fromloc.city_name	O	B-airline_name
FR	Montrer	des	départs	d’	Atlanta	pour	American
	O	O	O	O	B-fromloc.city_name	O	B-airline_name
ZH	显示	美国航空	从	亚特兰大	出发的航班		
	O	B-airline_name	O	B-fromloc.city_name	O		
JA	アトランタ	発	アメリカ	便を表示する			
	B-fromloc.city_name	O	B-airline_name	O			
HI	अमेरिकन	के	लिए	अटलांटा	से	प्रस्थान	दिखाएं
	B-airline_name	O	O	B-fromloc.city_name	O	O	O
TR	atlanta	‘	dan	american	kalkislarini	goster	
	B-fromloc.city_name	O	O	B-airline_name	O	O	

Figure 1: An English training example and its translated versions in the MultiATIS++ corpus. The English utterance is manually translated to the other eight languages including Spanish (ES), Portuguese (PT), German (DE), French (FR), Chinese (ZH), Japanese (JA), Hindi (HI), and Turkish (TR). For each language, we show the utterance followed by the slot labels in the BIO format. The intent of the utterances is the *flight* intent.

Intent acc.		en	es	de	zh	ja	pt	fr	hi	tr
Target only	LSTM	96.08	93.04	94.02	92.50	91.18	92.70	94.71	84.46	81.12
	BERT	**97.20**	**96.44**	**96.73**	**95.52**	95.54	**96.71**	**97.38**	90.50	87.10
Multilingual	LSTM	95.45	94.09	95.05	93.42	92.90	94.02	94.80	87.79	85.43
	BERT	**97.20**	**96.77**	**96.86**	**95.54**	**96.44**	96.48	**97.24**	**92.70**	**92.20**
Slot F1		en	es	de	zh	ja	pt	fr	hi	tr
Target only	LSTM	94.71	75.89	91.44	90.84	88.80	88.43	85.93	74.93	64.43
	BERT	95.57	86.58	**94.98**	**93.52**	91.40	91.35	89.14	82.36	75.21
Multilingual	LSTM	94.75	84.11	92.00	90.76	88.55	88.79	87.96	77.34	77.25
	BERT	**95.90**	**87.95**	**95.00**	**93.67**	**92.04**	**91.96**	**90.39**	**86.73**	**86.04**

Table 2: Results on MultiATIS++ using full training data and the standard supervised objective averaged over 5 runs. The *Target only* models are trained only on the target language training data. The *Multilingual* models are trained on the concatenation of training data from all languages.

SNIPS

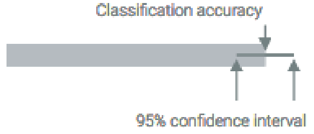
- 2016
- Very focused on privacy (their company value add?)
- 328 Test queries built by business team at SNIPS
- Unclear if they just used APIs from other companies

SNIPS

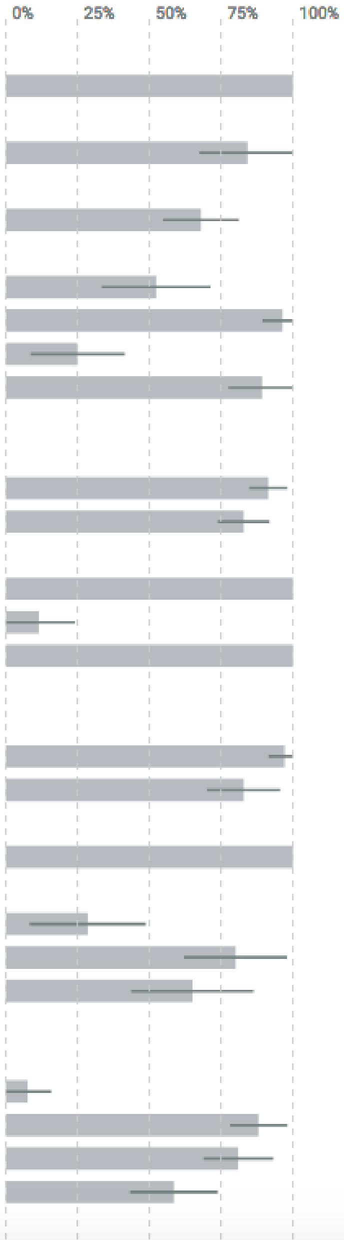
‘‘Built-in intents are compared to similar domains across providers. Providers not offering built-in intents were not included in this benchmark. The performance of these service depends on the efforts made by developers when training their custom intents, which makes comparisons methodologically more complicated. In the end, we tested the following services: Snips (10 intents), Google Api.ai (6 intents), Amazon Alexa (2 intents), Microsoft Luis (3 intents) and Apple Siri (1 intent).”

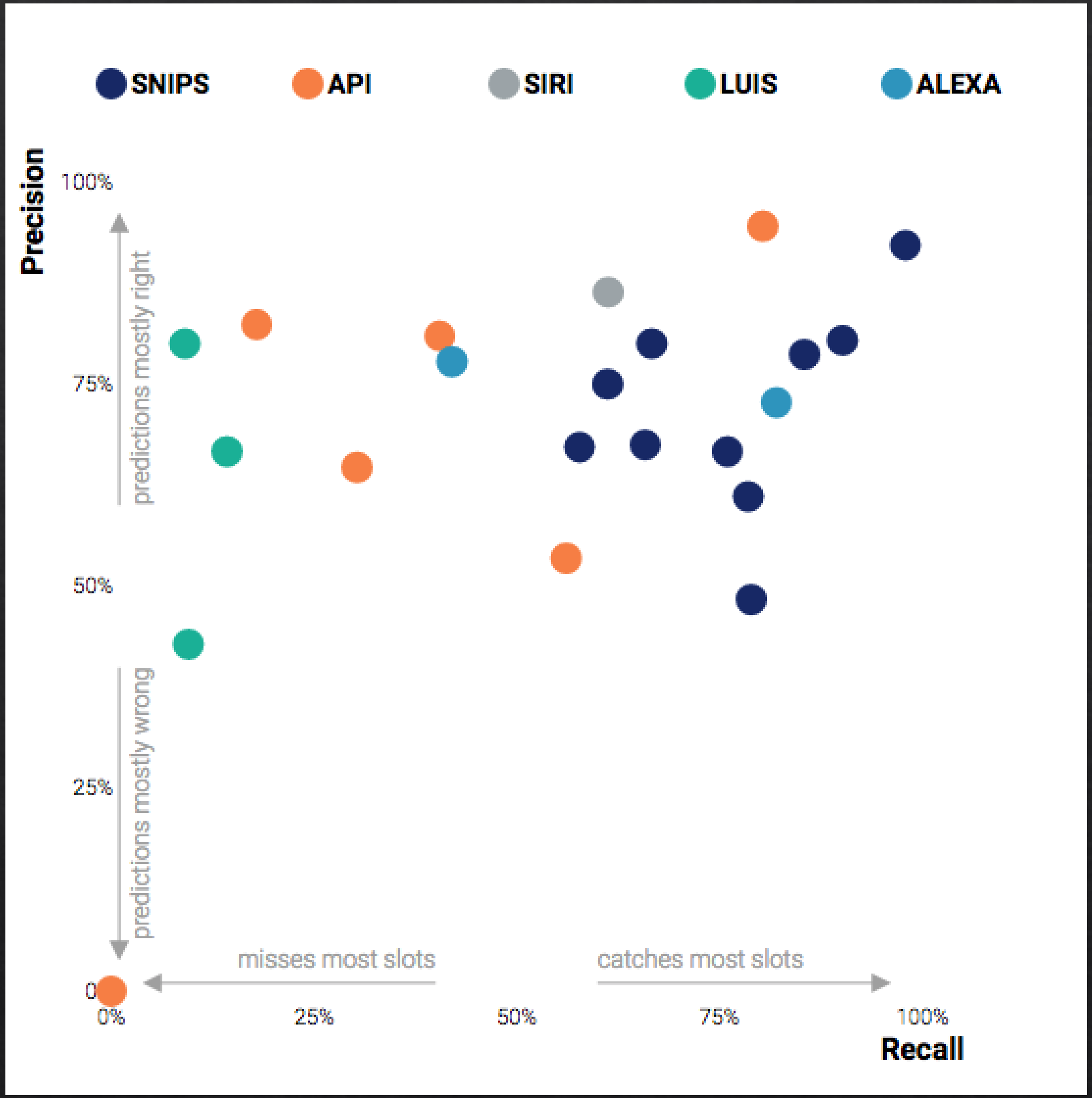
```
{
    "domains": [
        {
            "description": "Queries that are related to places (restaurants, shops, concert halls, etc), as well as to the user's location.",
            "@type": "domain",
            "intents": [
                {
                    "description": "The user wants to share his/her current location with someone.",
                    "benchmark": {
                        "Snips": {
                            "f1": 0.7272727272727272,
                            "classification_accuracy_2std": 0.0,
                            "n_queries": 16,
                            "classification_accuracy": 1.0,
                            "precision": 0.8,
                            "original_intent_name": "ShareCurrentLocation",
                            "slots": [
                                {
                                    "f1": 0.88,
                                    "description": "The person the user wants to send his/her location to.",
                                    "n_queries": 11,
                                    "precision": 0.7857142857142857,
                                    "slot_entity": "Contact",
                                    "name": "contact",
                                    "recall": 1.0,
                                    "precision_2std": 0.24743582965269675,
                                    "matching_slots": [
                                        {
                                            "slot": "contact",
                                            "service": "Snips"
                                        }
                                    ],
                                    "recall_2std": 0.0,
                                    "f1_2std": 0.19595917942265423
                                },
                                {
                                    "f1": 0.25,
                                    "description": "The length of the period over which the user wants to share his/her location.",
                                    "n_queries": 7,
                                    "precision": 1.0,
                                    "slot_entity": "Duration",
                                    "name": "sharingDuration",
                                    "recall": 0.14285714285714285,
                                    "precision_2std": 0.0,
                                    "matching_slots": [
                                        {
                                            "slot": "sharingDuration",
                                            "service": "Snips"
                                        }
                                    ],
                                    "recall_2std": 0.26452002850644329,
                                    "f1_2std": 0.32732683535398854
                                }
                            ],
```

ABOUT THE CHART



		CLASSIFICATION ACCURACY:	
PLACES			
ShareCurrentLocation	SNIPS	100%	(+/- 0%)
ComparePlaces	SNIPS	84%	(+/- 17%)
GetPlaceDetails	SNIPS	68%	(+/- 13%)
places.find_place	LUIS	52%	(+/- 19%)
SearchPlace	SNIPS	96%	(+/- 7%)
maps.places	API	25%	(+/- 16%)
LocalBusiness	ALEXA	89%	(+/- 12%)
RESERVATION			
BookRestaurant	SNIPS	91%	(+/- 7%)
book.restaurant	API	83%	(+/- 9%)
RequestRide	SNIPS	100%	(+/- 0%)
taxi.search	API	12%	(+/- 13%)
RequestRideIntent	SIRI	100%	(+/- 0%)
TRANSIT			
GetDirections	SNIPS	97%	(+/- 6%)
maps.directions	API	83%	(+/- 13%)
ShareETA	SNIPS	100%	(+/- 0%)
places.check_area_traffic	LUIS	29%	(+/- 20%)
GetTrafficInformation	SNIPS	80%	(+/- 18%)
maps.traffic	API	65%	(+/- 21%)
WEATHER			
weather.question_weather	LUIS	8%	(+/- 8%)
GetWeather	SNIPS	88%	(+/- 10%)
weather.search	API	81%	(+/- 12%)
WeatherForecast	ALEXA	59%	(+/- 15%)




SNIPS
API
SIRI
LUIS
ALEXA
Precision
100%
75%
50%
25%
0
predictions mostly right
predictions mostly wrong
misses most slots
catches most slots
0%
25%
50%
75%
100%
Recall

1. Filter queries per domain and intent

Domain: ALL DOMAINS | PLACES | RESERVATION | TRANSIT | WEATHER

Intent : All intents | ShareCurrentLocation | ComparePlaces | GetPlaceDetails | SearchPlace | BookRestaurant | RequestRide | GetDirections | ShareETA | GetTrafficInformation | GetWeather

2. Navigate through queries

Book a table for friday 8pm for 2 people at Katz's Delicatessen

Previous Next

3. Vizualise predictions

SNIPS

PREDICTED INTENT: **BookRestaurantQuery** ✔

SLOT		VALUE
reservationDatetime	✔	*2016-12-23T20:00*
partySize	✔	*2*
restaurant	✔	*Katzs Delicatessen*

API

PREDICTED INTENT: **book.restaurant**

SLOT		VALUE
people	✔	*2.0*
time	✔	*20:00:00*
date	✔	*2016-12-23*
venue_title	✖	*Delicatessen*

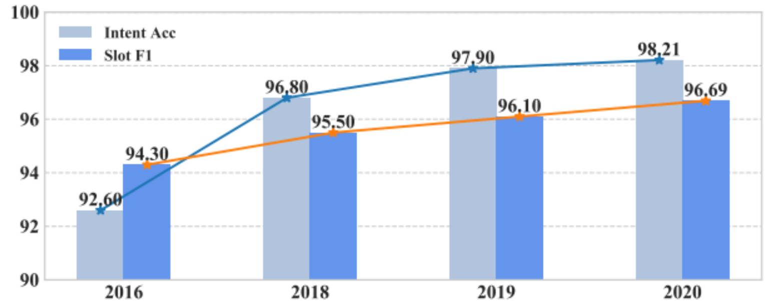
Evaluation

- F1 Scores
  - slot filling
  - exact match
- Intent Accuracy
  - intent detection
  - ratio of sentences correct
- Overall Accuracy
  - both intent, and slot predicted correctly
  - ratio of sentences correct

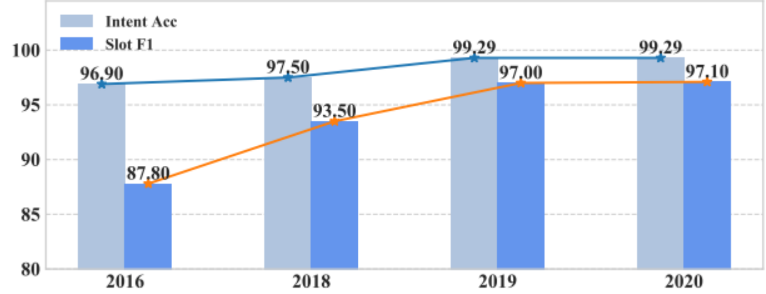
Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)
Survey Track

A Survey on Spoken Language Understanding: Recent Advances and New Frontiers

Libo Qin, Tianbao Xie, Wanxiang Che*, Ting Liu
Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology, China
{lbqin, tianbaoxie, car, tliu}@ir.hit.edu.cn



(a) Performance Trend on ATIS.



(b) Performance Trend on SNIPS.

Figure 2: Recent Performance Trend.

Single Models

Model	Intent Acc	Slot F1
Bi-Jordan RNN [Mesnil *et al.*, 2013]	-	93.98
RNN [Yao *et al.*, 2013]	-	94.11
Hybrid RNN [Mesnil *et al.*, 2014]	-	95.06
LSTM [Yao *et al.*, 2014a]	-	95.08
R-CRF [Yao *et al.*, 2014b]	-	96.65
RNN [Ravuri and Stolcke, 2015]	97.55	-
LSTM [Ravuri and Stolcke, 2015]	98.06	-
RNN SOP [Liu and Lane, 2015]	-	94.89
5xR-biRNN [Vu *et al.*, 2016]	-	95.56
Encoder-labeler [Kurata *et al.*, 2016]	-	95.66

Table 1: Single model performance on intent detection and slot filling on ATIS. Acc denotes the accuracy metric.

Joint Modeling (SOTA?)

BERT for Joint Intent Classification and Slot Filling

Qian Chen*, Zhu Zhuo, Wen Wang
Speech Lab, DAMO Academy, Alibaba Group
{tanqing.cq, zhuozhu.zz, w.wang}@alibaba-inc.com

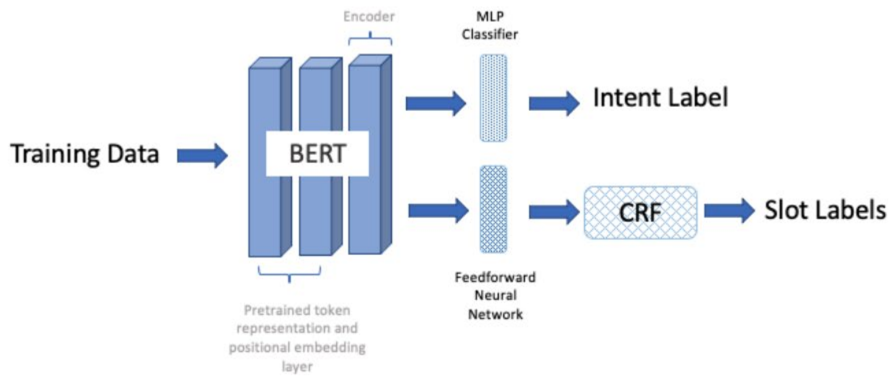


Figure 1: IC/SF Bert architecture

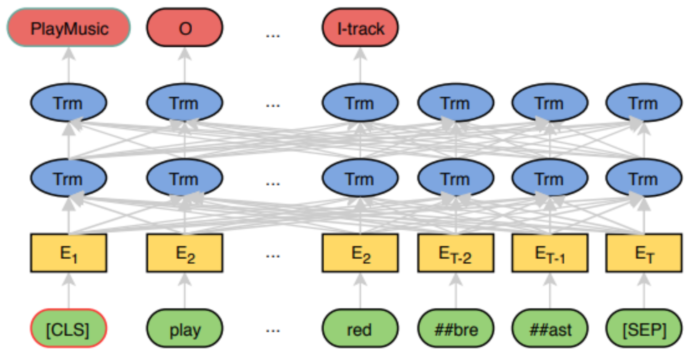


Figure 1: A high-level view of the proposed model. The input query is "play the song little robin redbreast".

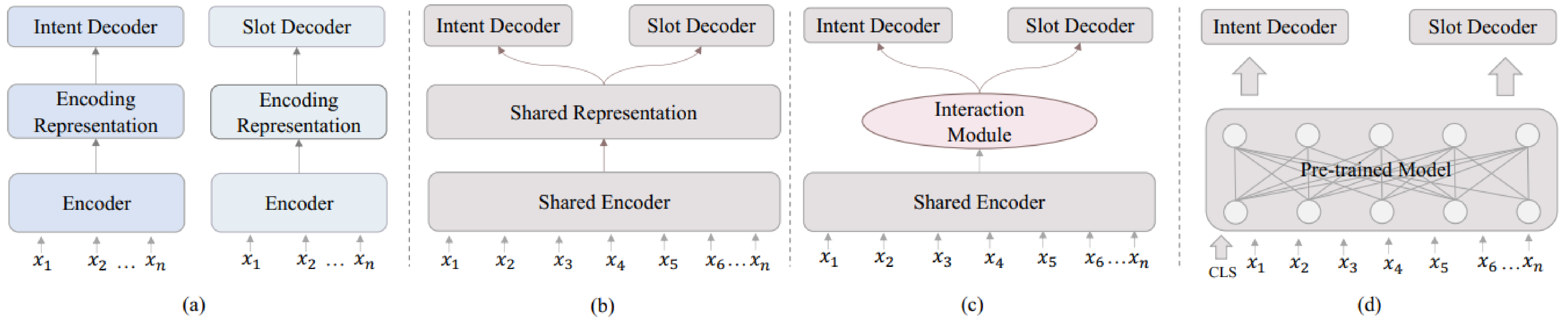


Figure 3: (a) Single models. (b) Implicit Joint Modeling. (c) Explicit Joint Modeling. (d) Pre-trained Model Paradigm.

	ATIS			SNIPS		
Model	Intent Acc	Slot F1	Overall Acc	Intent Acc	Slot F1	Overall Acc
Implicit Joint Modeling						
Joint ID and SF [Zhang and Wang, 2016]	98.32	96.89	-	-	-	-
Attention BiRNN [Liu and Lane, 2016a]	91.10	94.20	78.90	96.70	87.80	74.10
Joint SLU-LM [Liu and Lane, 2016b]	98.43	94.47	-	-	-	-
Joint Seq. [Hakkani-Tür *et al.*, 2016]	92.60	94.30	80.70	96.90	87.30	73.20
Explicit Joint Modeling						
Slot-Gated [Goo *et al.*, 2018]	93.60	94.80	82.20	97.00	88.80	75.50
Self-Atten. Model [Li *et al.*, 2018]	96.80	95.10	82.20	97.50	90.00	81.00
Bi-model [Wang *et al.*, 2018]	96.40	95.50	85.70	97.20	93.50	83.80
SF-ID Network [E *et al.*, 2019]	97.09	95.80	86.90	97.29	92.23	80.43
Capsule-NLU [Zhang *et al.*, 2019]	95.00	95.20	83.40	97.30	91.80	80.90
CM-Net [Liu *et al.*, 2019]	96.10	95.60	85.30	98.00	93.40	84.10
Stack-Propgation [Qin *et al.*, 2019]	96.90	95.90	86.50	98.00	94.20	86.90
Graph LSTM [Zhang *et al.*, 2020b]	97.20	95.91	87.57	98.29	95.30	89.71
Co-Interactive transformer [Qin *et al.*, 2021b]	97.70	95.90	87.40	98.80	95.90	90.30
Pre-trained Models						
BERT-Joint [Castellucci *et al.*, 2019]	97.80	95.70	88.20	99.00	96.20	91.60
Joint BERT +CRF [Chen *et al.*, 2019]	97.90	96.00	88.60	98.40	96.70	92.60
Stack-Propgation +BERT [Qin *et al.*, 2019]	97.50	96.10	88.60	99.00	97.00	92.90
Co-Interactive transformer +BERT [Qin *et al.*, 2021b]	98.00	96.10	88.80	98.80	97.10	93.10

Table 2: Joint model performance on intent detection and slot filling. Acc denotes the accuracy metric. We adopted reported results from published literature [Goo *et al.*, 2018] and [Qin *et al.*, 2021b].

Cross-Lingual SLU

- 2020
- Multi-ATIS++
- 9 languages

End-to-End Slot Alignment and Recognition for Cross-Lingual NLU

Weijia Xu*
University of Maryland
weijia@cs.umd.edu

Batool Haider
Amazon AI
bhaider@amazon.com

Saab Mansour
Amazon AI
saabm@amazon.com

Cross-Lingual SLU

- 2019
- English (43k)
- Sample En → Spanish (8.6k)
- Sample En → Thai (5k)
- Domains {Weather, Alarm, Reminder}

Cross-Lingual Transfer Learning for Multilingual Task Oriented Dialog

Sebastian Schuster*
Stanford Linguistics
sebschu@stanford.edu

Sonal Gupta
Facebook Conversational AI
sonalgupta@fb.com

Rushin Shah
Facebook Conversational AI
rushinshah@fb.com

Mike Lewis
Facebook AI Research
mikelewis@fb.com

Domain	Number of utterances English	Spanish	Thai	Intent types	Slot types
Alarm	9,282/1,309/2,621	1,184/691/1,011	777/439/597	6	2
Reminder	6,900/943/1,960	1,207/647/1,005	578/336/442	3	6
Weather	14,339/1,929/4,040	1,226/645/1,027	801/460/653	3	5
Total	30,521/4,181/8,621	3,617/1,983/3,043	2,156/1,235/1,692	12	11

Table 1: Summary statistics of the data set. The three values for the number of utterances correspond to the number of utterances in the training, development, and test splits. Note that the slot type *datetime* is shared across all three domains and therefore the total number of slot types is only 11.

Cross-Lingual SLU

- 2020
- Code-Switching
- Uses Schuster et al., 2019
- Bilingual Dictionary

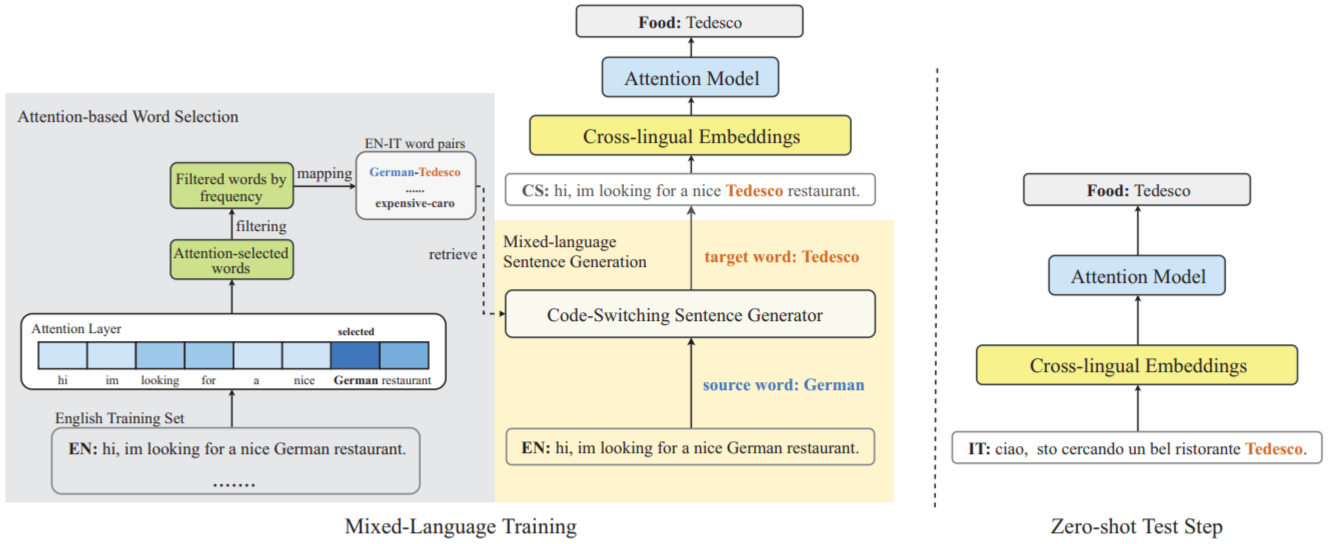


Figure 1: Illustration of the mixed-language training (MLT) approach and zero-shot transfer. **EN** denotes an English text, **IT** denotes an Italian text, and **CS** denotes a code-switching text (i.e., a mixed-language sentence). In the training step, code-switching sentence generator will replace the task-related word with its corresponding translation in the target language to generate code-switching sentences. In the zero-shot transfer step, we leverage cross-lingual word embeddings and directly adapt the trained attention model to the target language.

The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)

Attention-Informed Mixed-Language Training for Zero-Shot Cross-Lingual Task-Oriented Dialogue Systems

Zihan Liu,* Genta Indra Winata,* Zhaojiang Lin, Peng Xu, Pascale Fung
Center for Artificial Intelligence Research (CAiRE)
The Hong Kong University of Science and Technology
{zliucr, giwinata, zlinao, pxuab}@connect.ust.hk, pascale@ece.ust.hk

- Multilingual WOZ 2.0, Mrkšić et al. 2017b
- Restaurant Domain, Slot Filling (Dialogue State Tracking)
- Translated into German and Italian

	German								
Model	slot acc.			joint goal acc.			request acc.		
	BASE	**MLT$_O$**	**MLT$_A$**	**BASE**	**MLT$_O$**	**MLT$_A$**	**BASE**	**MLT$_O$**	**MLT$_A$**
MUSE	60.69	68.58	**71.38**	21.57	30.61	**36.51**	74.22	80.11	**82.99**
XLM (MLM)*	52.21	66.26	**68.25**	14.09	29.45	**31.29**	75.15	78.48	**80.22**
+ Transformer	53.81	65.81	**68.55**	13.97	30.87	**32.98**	76.83	78.95	**81.34**
XLM (MLM+TLM)*	58.04	65.39	**66.25**	16.34	29.22	**29.83**	75.73	78.86	**79.12**
+ Transformer	56.52	66.81	**68.88**	16.59	31.76	**33.12**	78.56	81.59	**82.96**
Multi. BERT*	57.61	67.49	**69.48**	14.95	30.69	**32.23**	75.31	83.66	**86.27**
+ Transformer	57.43	68.33	**70.77**	15.67	31.28	**34.36**	78.59	84.37	**86.97**
Ontology Matching†	*24*			-			*21*		
Translate Train†	*41*			-			*42*		
Bilingual Dictionary‡	*51.74*			*28.07*			*72.54*		
Bilingual Corpus‡	*55*			*30.84*			*68.32*		
Supervised Training	*85.78*			*78.89*			*84.02*		

	Italian								
Model	slot acc.			joint goal acc.			request acc.		
	BASE	**MLT$_O$**	**MLT$_A$**	**BASE**	**MLT$_O$**	**MLT$_A$**	**BASE**	**MLT$_O$**	**MLT$_A$**
MUSE	60.59	73.55	**76.88**	20.66	36.88	**39.35**	79.09	82.24	**84.23**
Multi. BERT*	53.34	65.49	**69.48**	12.88	26.45	**31.41**	76.12	84.58	**85.18**
+ Transformer	54.56	66.87	**71.45**	12.63	28.59	**33.35**	77.34	82.93	**84.96**
Ontology Matching†	*23*			-			*21*		
Translate Train†	*48*			-			*51*		
Bilingual Dictionary‡	*73*			*39.01*			*77.09*		
Bilingual Corpus‡	*72*			*41.23*			*81.23*		
Supervised Training	*88.92*			*80.22*			*91.05*		

Table 1: Zero-shot results for the target languages on **Multilingual WOZ 2.0**. **MLT$_A$** denotes our approach (attention-informed MLT), which utilizes the same number of word pairs as **MLT$_O$** (90 word pairs). ‡ denotes the results of XL-NBT. Note that, we realize that the goal accuracy in Chen et al. (2018) is calculated as slot accuracy in our paper, so we rerun the models using the provided code (https://github.com/wenhuchen/Cross-Lingual-NBT) to calculate joint goal accuracy. † denotes the results from Chen et al. (2018). Instead of using the *transformer encoder*, we sum the subword embeddings based on the word boundaries to get word-level representations. Due to the absence of the Italian language in the XLM models, we cannot report the results.

Cross-Lingual SLU

- 2020
- Code-Switching
- Similar authors
- Fine-tune mBERT on random word code-mixing data (CLCSA)

Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)

CoSDA-ML: Multi-Lingual Code-Switching Data Augmentation for Zero-Shot Cross-Lingual NLP

Libo Qin[1], **Minheng Ni**[1], **Yue Zhang**[2,3], **Wanxiang Che**[1]
[1]Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology, China
[2]School of Engineering, Westlake University, China
[3]Institute of Advanced Technology, Westlake Institute for Advanced Study
{lbqin, mhni, car}@ir.hit.edu.cn, yue.zhang@wias.org.cn

Model	German			Italian		
	slot acc.	joint goal acc.	request acc.	slot acc.	joint goal acc.	request acc.
XL-NBT [Chen *et al.*, 2018]	55.0	30.8	68.4	72.0	41.2	81.2
Attention-Informed Mixed Training [Liu *et al.*, 2019b]	69.5	32.2	86.3	69.5	31.4	85.2
XLM from Liu *et al.* [2019b]	58.0	16.3	75.7	-	-	-
+CLCSA	77.4	48.7	88.3	-	-	-
mBERT [Devlin *et al.*, 2019]	57.6	15.0	75.3	54.6	12.6	77.3
+CLCSA	**83.0***	**63.2***	**94.0***	**82.2***	**61.3***	**94.2***

Table 4: Dialog State Tracking experiments.

Model	Spanish		Thai	
	Intent acc.	Slot F1	Intent acc.	Slot F1
Multi. CoVe [Yu *et al.*, 2018]	53.9	19.3	70.7	35.6
Attention-Informed Mixed Training [Liu *et al.*, 2019b]	86.5	74.4	70.6	28.5
XLM from Liu *et al.* [2019b]	62.3	42.3	31.6	7.9
+ CLCSA	90.3	69.0	**86.7**	34.9
mBERT [Devlin *et al.*, 2019]	73.7	51.7	28.2	10.6
+ CLCSA (Static)	92.8	75.2	74.8	28.1
+ CLCSA	**94.8***	**80.4***	76.8	**37.3***

Table 5: Slot filling and Intent detection experiments.

Chinese SLU (CAIS)

- 2019
- Really underspecified

CM-Net: A Novel Collaborative Memory Network for Spoken Language Understanding

Yijin Liu[1*], Fandong Meng[2], Jinchao Zhang[2], Jie Zhou[2], Yufeng Chen[1] and Jinan Xu[1†]
[1]Beijing Jiaotong University, China
[2]Pattern Recognition Center, WeChat AI, Tencent Inc, China
adaxry@gmail.com
{fandongmeng, dayerzhang, withtomzhou}@tencent.com
{chenyf, jaxu}@bjtu.edu.cn

MASSIVE

- Amazon
- 2022
- 51 Languages
- 18 Domains
- 60 Intents
- 55 Slots
- 29 Genera
- SLURP Dataset → Localize/Translate (Professionals)

MASSIVE: A 1M-Example Multilingual Natural Language Understanding Dataset with 51 Typologically-Diverse Languages

Jack FitzGerald* **Christopher Hench** **Charith Peris**
Scott Mackie **Kay Rottmann** **Ana Sanchez**
Aaron Nash **Liam Urbach** **Vishesh Kakarala**
Richa Singh **Swetha Ranganath** **Laurie Crist**
Misha Britan **Wouter Leeuwis** **Gokhan Tur**
Prem Natarajan

Name	# Lang	Utt per Lang	Domains	Intents	Slots
MASSIVE	51	19,521	18	60	55
SLURP (Bastianelli et al., 2020)	1	16,521	18	60	55
NLU Evaluation Data (Liu et al., 2019a)	1	25,716	18	54	56
Airline Travel Information System (ATIS) (Price, 1990)	1	5,871	1	26	129
ATIS with Hindi and Turkish (Upadhyay et al., 2018)	3	1,315-5,871	1	26	129
MultiATIS++ (Xu et al., 2020)	9	1,422-5,897	1	21-26	99-140
Snips (Coucke et al., 2018)	1	14,484	-	7	53
Snips with French (Saade et al., 2019)	2	4,818	2	14-15	11-12
Task Oriented Parsing (TOP) (Gupta et al., 2018)	1	44,873	2	25	36
Multilingual Task-Oriented Semantic Parsing (MTOP) (Li et al., 2021)	6	15,195-22,288	11	104-113	72-75
Cross-lingual Multilingual Task Oriented Dialog (Schuster et al., 2019)	3	5,083-43,323	3	12	11
Microsoft Dialog Challenge (Li et al., 2018b)	1	38,276	3	11	29
Fluent Speech Commands (FSC) (Lugosch et al., 2019)	1	30,043	-	31	-
Chinese Audio-Textual Spoken Language Understanding (CATSLU) (Zhu et al., 2019)	1	16,258	4	-	94

Table 1: Selected NLU benchmark datasets with number of languages, utterances per language, domain count, intent count, and slot count.

Massive Adjacent (MMNLU-22)

HIT-SCIR at MMNLU-22: Consistency Regularization for Multilingual Spoken Language Understanding

Bo Zheng, Zhouyang Li, Fuxuan Wei, Qiguang Chen, Libo Qin, Wanxiang Che*
Harbin Institute of Technology
{bzheng,zhouyangli,fxwei,qgchen,lbqin,car}@ir.hit.edu.cn

Evaluating Byte and Wordpiece Level Models for Massively Multilingual Semantic Parsing

Massimo Nicosia and Francesco Piccinno
Google Research, Zürich
{massimon,piccinno}@google.com

Zero-Shot

Rank	Participant team	Exact Match Acc (↑)	Intent Acc (↑)	Slot Micro F1 (↑)	High Lang EMA (↑)	High Lang EMA Val (↑)	Low Lang EMA (↑)	Low Lang EMA Val (↑)
1	HIT-SCIR (Base Encoder Only + xTune)	44.84	83.89	64.60		50.53		35.57
2	FabT5 (Translate-and-Fill + ByT5)	44.43	82.71	64.00		51.20		37.60

HIT-SCIR at MMNLU-22: Consistency Regularization for Multilingual Spoken Language Understanding

Bo Zheng, Zhouyang Li, Fuxuan Wei, Qiguang Chen, Libo Qin, Wanxiang Che*
Harbin Institute of Technology
{bzheng,zhouyangli,fxwei,qgchen,lbqin,car}@ir.hit.edu.cn

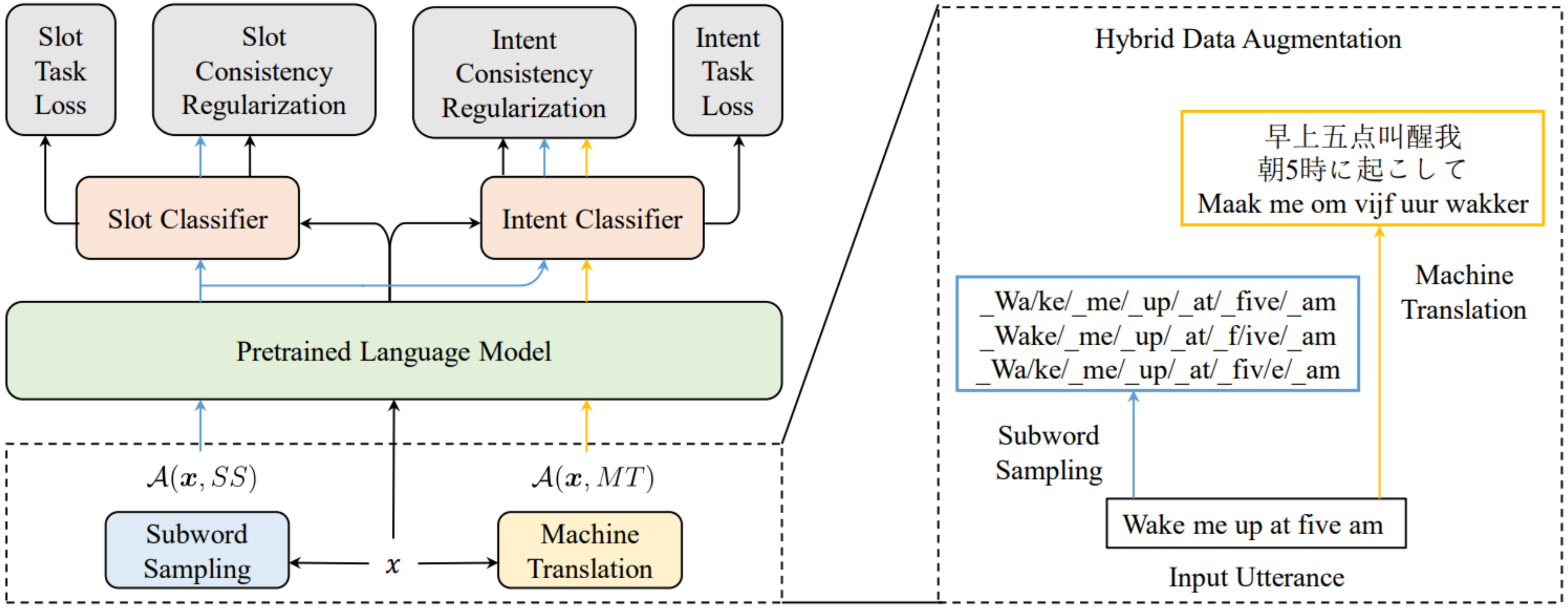


Figure 2: Illustration of our fine-tuning framework. 'MT' denotes machine translation augmentation and 'SS' denotes subword sampling augmentation.

Translate & Fill: Improving Zero-Shot Multilingual Semantic Parsing with Synthetic Data

Massimo Nicosia, Zhongdi Qu, Yasemin Altun

Google Research
{massimon,dqu,altun}@google.com

(a) Parser training instance

alarm at 8 am → [IN:CREATE_ALARM [SL:DATE_TIME 8 am]]

(b) Filler training instance

alarm at 8 am | [IN:CREATE_ALARM [SL:DATE_TIME]] → [IN:CREATE_ALARM [SL:DATE_TIME 8 am]]

(c) Filler inference input

sveglia alle 8 di mattina | [IN:CREATE_ALARM [SL:DATE_TIME]]

→

(d) Filler inference output

[IN:CREATE_ALARM [SL:DATE_TIME 8 di mattina]]

(e) Parser synthetic training instance

sveglia alle 8 di mattina → [IN:CREATE_ALARM [SL:DATE_TIME 8 di mattina]]

Figure 1: Example instances for training the semantic **parser** (a) and the **filler** (b). The filler is trained to produce a full parse from the concatenation of an English utterance and the corresponding parse signature (b). At inference, we replace the English utterance with its (Italian in this case) translation (c), and obtain a silver parse where the slots contain text from the translation (d). The latter is used to assemble a synthetic training instance (e) for a multilingual semantic parser.

Translate & Fill: Improving Zero-Shot Multilingual Semantic Parsing with Synthetic Data

Massimo Nicosia, Zhongdi Qu, Yasemin Altun

Google Research
{massimon,dqu,altun}@google.com

Translate & Fill

Evaluating Byte and Wordpiece Level Models for Massively Multilingual Semantic Parsing

Massimo Nicosia and Francesco Piccinno
Google Research, Zürich
{massimon,piccinno}@google.com

FabT5

FabT5

Intent	IA	Support
GENERAL_GREET	19.6	51
MUSIC_SETTINGS	27.1	306
AUDIO_VOLUME_OTHER	54.9	306
GENERAL_QUIRKY	55.6	8619
IOT_HUE_LIGHTON	61.4	153
MUSIC_DISLIKENESS	74.5	204

Table 5: IA of the ByT5-xxl+TAF model for the lowest scoring intents (considering all languages).

FabT5

- This may suggest that translations in these languages are more unambiguous or that translators may have relied on a MT during the translation task.

Language sets	**Avg Match** (%)
All languages	21.3
All but Indic languages	17.3
Indic languages	50.8

Table 6: Percentages of NMT translations matching human translations in MASSIVE training set.

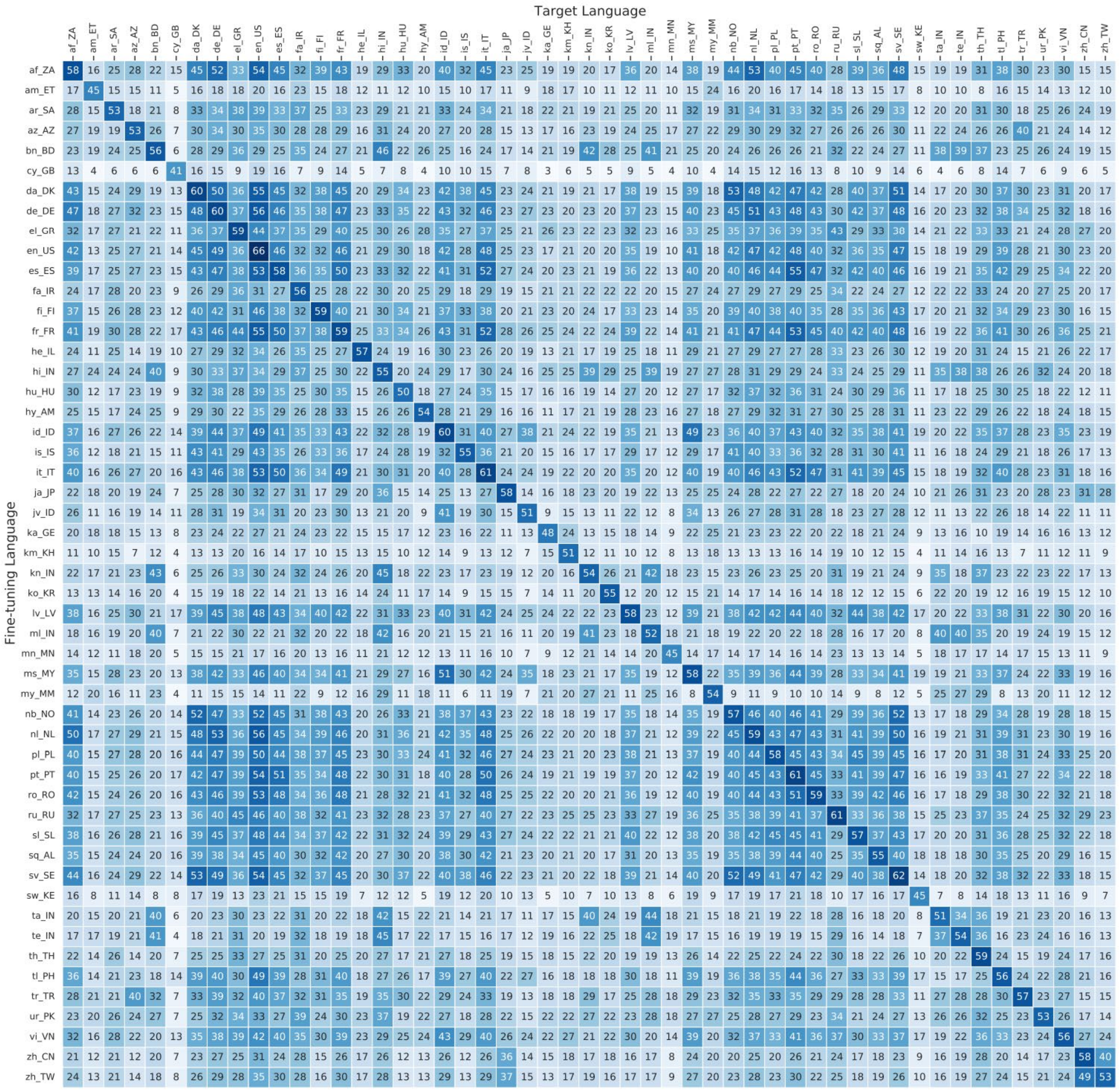


Figure 2: Zero-shot EM accuracies of individual ByT5-base models fine-tuned on a single language (y-axis) and evaluated on dev sets from all languages (x-axis).

◈ On many of the important papers/datasets

Gokhan Tur

FOLLOW

Amazon Alexa AI
Verified email at ieee.org

Conversational AI   Language Understanding   Deep Learning   Machine Learning
Statistical Speech and Lan...

Cited by   VIEW ALL

	All	Since 2017
Citations	13392	8172
h-index	56	45
i10-index	176	125

Dilek Hakkani-Tur

FOLLOW

Amazon Alexa AI
Verified email at ieee.org - Homepage

Speech and Language Pro...   Dialogue Systems   Spoken Language Underst...
Machine Learning   Dialog Management

Cited by   VIEW ALL

	All	Since 2018
Citations	17901	9915
h-index	71	51
i10-index	280	176

Larry Heck

FOLLOW

Professor, Georgia Institute of Technology
Verified email at ieee.org - Homepage

dialogue   conversational AI   natural language understan...   natural language generation
spoken language processing

Cited by   VIEW ALL

	All	Since 2018
Citations	10707	6959
h-index	51	39

Manaal Faruqui

FOLLOW

Research Scientist, Google
Verified email at google.com - Homepage

natural language processing

Cited by   VIEW ALL

	All	Since 2018
Citations	4944	3726
h-index	28	27
i10-index	36	34

Last Words

Revisiting the Boundary between ASR and NLU in the Age of Conversational Dialog Systems

Manaal Faruqui
Google Assistant
mfaruqui@google.com

Dilek Hakkani-Tür
Amazon Alexa AI
hakkanit@amazon.com

As more users across the world are interacting with dialog agents in their daily life, there is a need for better speech understanding that calls for renewed attention to the dynamics between research in automatic speech recognition (ASR) and natural language understanding (NLU). We briefly review these research areas and lay out the current relationship between them. In light of the observations we make in this paper, we argue that (1) NLU should be cognizant of the presence of ASR models being used upstream in a dialog system's pipeline, (2) ASR should be able to learn from errors found in NLU, (3) there is a need for end-to-end datasets that provide semantic annotations on spoken input, (4) there should be stronger collaboration between ASR and NLU research communities.

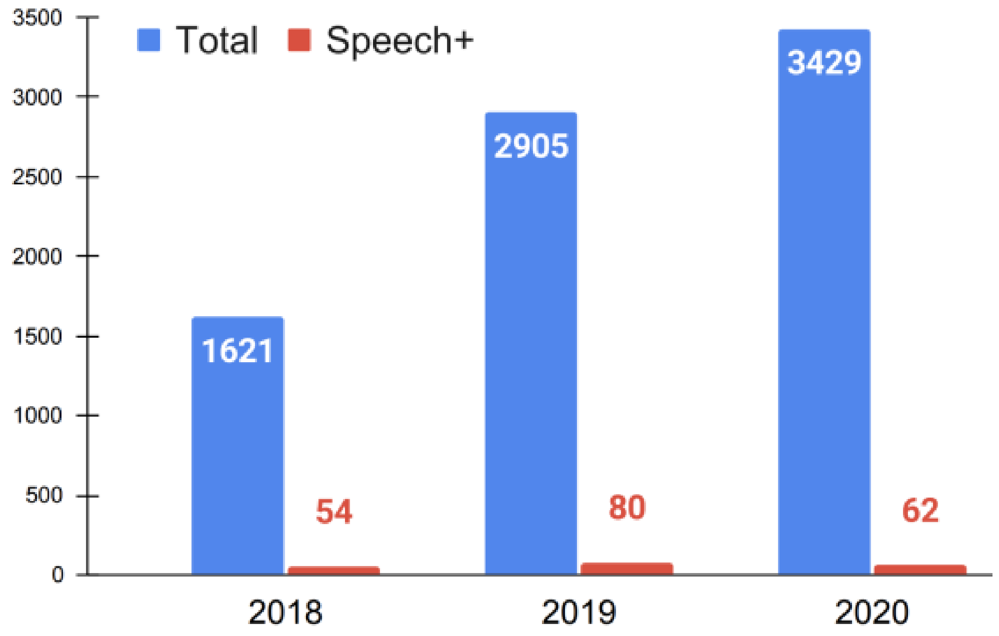


Figure 1
The number of submitted papers in the speech processing (+ multimodal) track vs. total in ACL conference from 2018-2020.

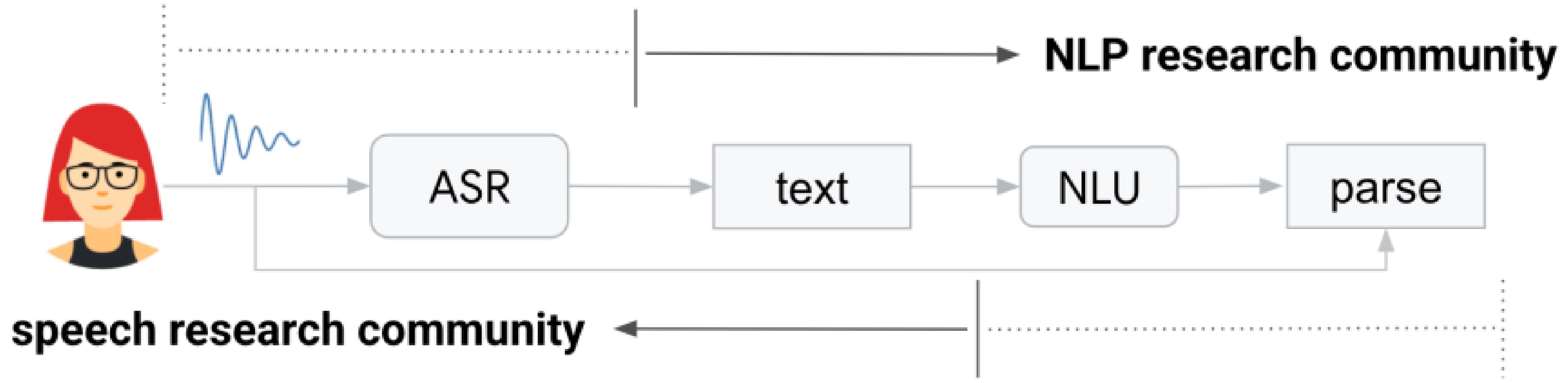


Figure 2
The current focus of speech and NLU research community (dark lines) and preferred focus of speech and NLU community (dotted lines) in future.

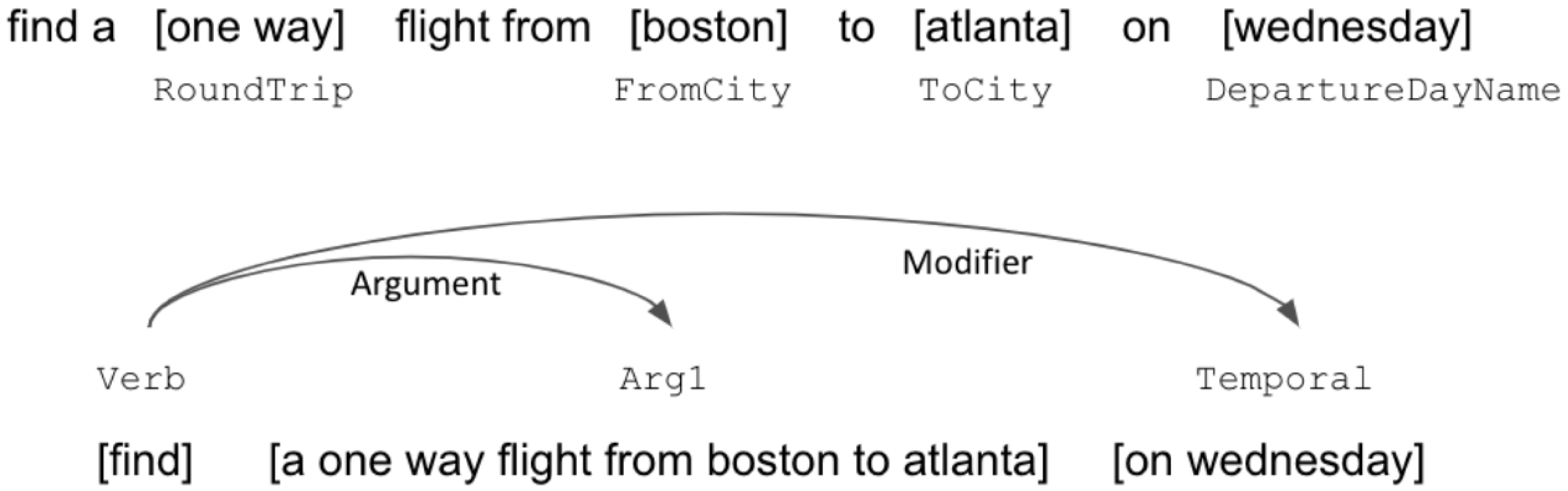


Figure 3
SLU annotation (top) and NLU semantic role labeling annotation (bottom) on a sentence from the English ATIS corpus (Price 1990), a popular SLU benchmark.