Developing a Production System for Purpose of Call Detection in Business Phone Conversations

Elena Khasanova, Pooja Hiranandani, Shayna Gardiner,
Cheng Chen, Xue-Yong Fu, Simon Corston-Oliver
Dialpad Canada Inc
1100 Melville St #400
Vancouver, BC, Canada, V6E 4A6
{elena.khasanova,phiranandani,sgardiner}@dialpad.com
{cchen,xue-yong,scorston-oliver}@dialpad.com

Abstract

For agents at a contact centre receiving calls, the most important piece of information is the reason for a given call. An agent cannot provide support on a call if they do not know why a customer is calling. In this paper we describe our implementation of a commercial system to detect *Purpose of Call* statements in English business call transcripts in real time. We present a detailed analysis of types of Purpose of Call statements and language patterns related to them, discuss an approach to collect rich training data by bootstrapping from a set of rules to a neural model, and describe a hybrid model which consists of a transformer-based classifier and a set of rules by leveraging insights from the analysis of call transcripts. The model achieved 88.6 F1 on average in various types of business calls when tested on real life data and has low inference time. We reflect on the challenges and design decisions when developing and deploying the system.

1 Introduction

The Purpose of Call as we define it is similar to a thesis statement in an argument: it introduces the speaker's intent, the broad theme or topic of a conversation, any key entities, and relevant relationships between them. The Purpose of Call statement might also include a linguistic signpost – an indication to the listener that the utterance is intended to convey the purpose of the speaker's call.

For instance:

I'm *calling because* [signpost] I'm trying to open *one of the programs* [entity] on *my computer* [entity] and *it's not opening* [relation] so I'm hoping I can *get some assistance* [intent] with that. [1]

Purpose of Call statements in a contact centre setting are usually uttered by the customer in inbound calls, and by the agent in outbound calls. The Purpose of Call is typically stated near the beginning of the call, is often stated in a single utterance, and does not contain extra information. Atypically, we may sometimes see the Purpose of Call occurring in the middle of a conversation, occurring across several utterances, being implicit, or being uttered by a call recipient rather than a call initiator.

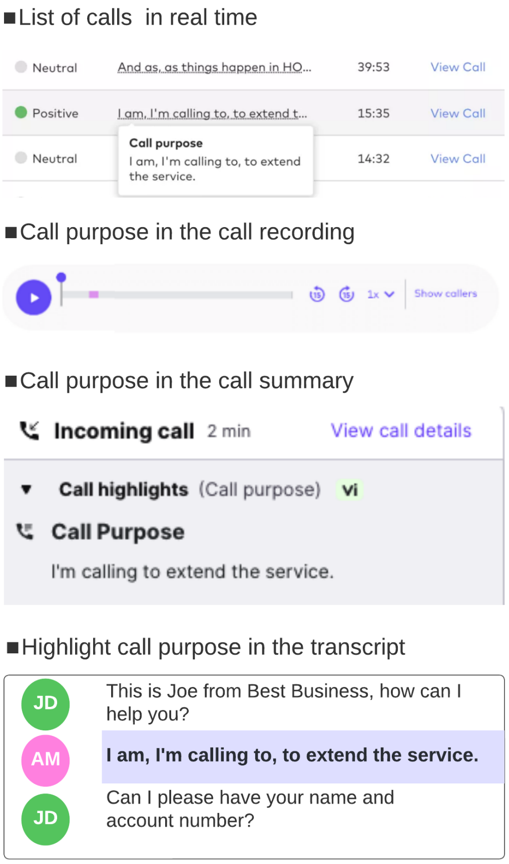
The models described below have been implemented in the Dialpad Contact Center product and are running in production. The Purpose of Call is extracted from an automatic speech recognition (ASR) generated transcript in near-realtime and displayed in a dashboard used by call center supervisors to monitor calls taking place. The dashboard shows information about the caller and the agent, the duration of the call, the Purpose of Call, and customer sentiment. A separate analytics component clusters the Purpose of Call from all calls in a call center during a time period to provide insights about trends and anomalies, customer pain points, and common problems and knowledge gaps among agents. Additional use-cases include showing the Purpose of Call in a summary of prior calls with a customer, and including the Purpose of Call in summaries of the conversation. The utterance segment containing the Purpose of Call is highlighted in the call transcript and the call recording to be easily accessible to agents and call center supervisors. These use-cases are summarized in Figure 1. The Purpose of Call feature is used to help call center managers to navigate to relevant sections of conversations to identify areas to coach sales and support agents and sample relevant calls. Through customer education, we emphasize that the feature should not be used for automated evaluation of agent performance.

There are a few challenges that arise when building an automatic system to detect Purpose of Call.

Diversity of Purpose of Call statements. This type of detection system should be an open-class

[1] In contrast, statements such as *I'm calling to ask a question* are not considered Purpose of Call expressions even though they contain relevant signposting language because there are no entities an agent or a customer can relate to.

Proceedings of NAACL-HLT 2022: Industry Track Papers, pages 259 - 267
July 10-15, 2022 ©2022 Association for Computational Linguistics


■List of calls in real time
■Call purpose in the call recording
1x
Show callers
■Call purpose in the call summary
View call details
■Highlight call purpose in the transcript
JD
AM
JD

Neutral	And as, as things happen in HO...	39:53	View Call
Positive	I am, I'm calling to, to extend t...	15:35	View Call
Neutral		14:32	View Call

Call purpose
I am, I'm calling to, to extend the service.

Incoming call 2 min

Call highlights (Call purpose) vi

Call Purpose

I'm calling to extend the service.

This is Joe from Best Business, how can I help you?

I am, I'm calling to, to extend the service.

Can I please have your name and account number?

Figure 1: An illustration of the applications of Purpose of Call

system, since call purposes vary across different domains, industries, and types of calls.

Robustness to noise. There are challenges related to the fact that Purpose of Call extraction relies on the output of an ASR system. There are two sources of noise in the ASR transcripts: language production issues such as false starts, dysfluencies, filled pauses, inconsistency in conversational turn-taking (Cailliau and Cavet, 2013; Dutrey et al., 2014; Żelasko et al., 2019; Clavel et al., 2013) as well as their representation in the ASR system and recognition errors due to acoustic noise.

Limitations in training data. Existing intent detection datasets do not reflect real world settings (e.g. they do not distinguish the Purpose of Call from other intent-like statements, and are limited to a subset of domains). Manually annotating data (e.g. using crowd-sourced annotators) raises privacy concerns since annotators must have access to a full conversation transcript in order to find the best Purpose of Call. Annotation is a complex task that requires highly trained annotators, and is expensive and time-consuming because annotators must consider the larger context of the conversation to make a judgment.

Computational efficiency. The need for the system to extract the Purpose of Call statement in real time imposes constraints on memory consumption, latency, and inference speed.

This paper describes an end-to-end system to extract a Purpose of Call statement from the transcript of a business telephone call. Our contributions are three-fold:

1. Data analysis: we present a detailed analysis of language patterns and other features involved in call purpose detection;

2. Data: we describe a process to overcome a lack of training data by bootstrapping a deep learning model from a knowledge-engineered model and discuss the heuristics for developing such a model;

3. System: we describe optimizations that were done in an online commercial system to identify Purpose of Call statements in near-realtime (within three seconds of an utterance being transcribed). To evaluate the effectiveness of our approach, we examine the actual output of our production system.

2 Related work

The concept of a Purpose of Call statement has its roots in the Conversation Analysis framework (Schegloff and Sacks, 1973; Sacks et al., 1974). Within this framework, which combines perspectives from Linguistics and Sociology, a conversation is understood to be composed of turn-taking utterances, with each "turn" being indicated via linguistic and paralinguistic cues. Conversational turns often form adjacency pairs such as question-answer pairs or offer-acceptance/refusal pairs. There are other key aspects of a conversation as well. For instance, a conversation is likely to end after a linguistic cue known as a "closing" is given; likewise, there is usually a linguistic indicator that a conversation is being initiated: an opening (Schegloff and Sacks, 1973; Sacks et al., 1974; Pomerantz and Fehr, 2011). Most work analyzing telephone conversation openings within the framework of Conversation Analysis has been conducted on English, but similar patterns have been observed in other languages, including German and Farsi (Taleghani-Nikazm, 2002).

Within a contact center environment, the Purpose of Call is, like openings and closings, an integral aspect of the conversation (i.e. call) between customer and support agent. We propose that a Purpose of Call is a particular conversational feature that is necessary in contact center calls, and is distinct from the call opening, the body of the call, and the call closing.

The first 120 seconds of a customer support call are predictive of that call's outcome (Takeuchi et al., 2007; Hall et al., 2014). The Purpose of Call statement typically occurs within this timeframe, so highlighting a Purpose of Call in real time could provide agents with additional support in meeting customer needs.

3 Methodology

We formulate the Purpose of Call detection task as a binary classification problem. Each call, after being transcribed by the ASR system, is represented as a sequence of utterances, which may consist of one or more sentences. The division into utterances is based on acoustic features such as silent pauses and the length of a speech fragment.

For a given utterance, we determine the probability that the utterance is the Purpose of Call statement for that particular call. We impose the following constraints on this task: (i) For a given call, there is only one most probable Purpose of Call. (ii) Only calls with two call sides (agent and customer) are considered, which excludes multiparty business conversations. (iii) The model should make a prediction as the call is ongoing and therefore will not have access to the full conversation.

Due to the lack of available annotated data representing the concept of Purpose of Call, we followed an iterative approach to develop the model, consisting of three steps: (1) Computational Linguists on-staff conducted extensive linguistic analysis of transcripts to identify the characteristics of Purpose of Call statements. (2) We then implemented a knowledge-engineered approach to identify these Purpose of Call statements. (3) We bootstrapped from the knowledge-engineered solution, using it to label training data for a transformer-based approach. We select a transformer-based model as it is the current state-of-the-art in sequence classification and is known to have better generalization power than rule-based models.

We evaluate the performance using F1, Precision, and Hit rate, i.e. the number of calls in which a Purpose of Call was detected out of all available calls. We measure Hit rate in calls at least 30 seconds long, based on the observation that shorter calls may not include any content (e.g. because the caller hung up before starting the conversation). The model is tested on an automatically obtained validation set that represents 10% (18K utterances) of the training data, a manually annotated gold test set of 13215 utterances from 909 calls, and unlabeled samples from 600 real-life calls.

3.1 System Overview

The production system to detect a Purpose of Call utterance is a hybrid model consisting of three parts (see Figure 2).

The Selection model, or outer model, inputs an utterance, the previous context of the conversation, and the probabilities of previously detected Purpose of Call events. It consists of two sets of rules: (1) empirically derived filters that determine whether an incoming utterance is a candidate for a Purpose of Call and should be processed by the inner model, a successful candidate is within 180 seconds and 30 utterances in the call, and is between 4 and 150 tokens long; (2) rules that combine and compare scores from the inner model and set various thresholds for different linguistic types of call purpose statements (i.e. the utterances containing signposting language typically receive higher scores than other types and need a higher bar). Every time a new utterance qualifies to be a Purpose of Call, it is dynamically updated in the user interface. (See Appendix B for an example of a Selection rule.)

The Scoring model, or inner model, is implemented as a multiclass classification model which performs inference on a single utterance. We fine-tuned a transformer-based model for classification on proprietary labeled data. The model assigns to an utterance probabilities of it being a *call purpose*, *question*, or *negative* (not a *call purpose* or a *question*). The *question* class represents *question_response* pattern (see Table 1) and is used to boost probabilities of utterances that would otherwise be of the *negative* class.

The Simplification model. The utterance with the highest score is stripped of information that is irrelevant to the purpose of the call (e.g. *greetings, pleasantries, introductions, technical problems*). It consists of a small set of common expressions (many of which are reused from the knowledge-engineered model) to exclude from utterances and reduces the length of Purpose of Call utterances by 7.8% on average. 49% of utterances undergo

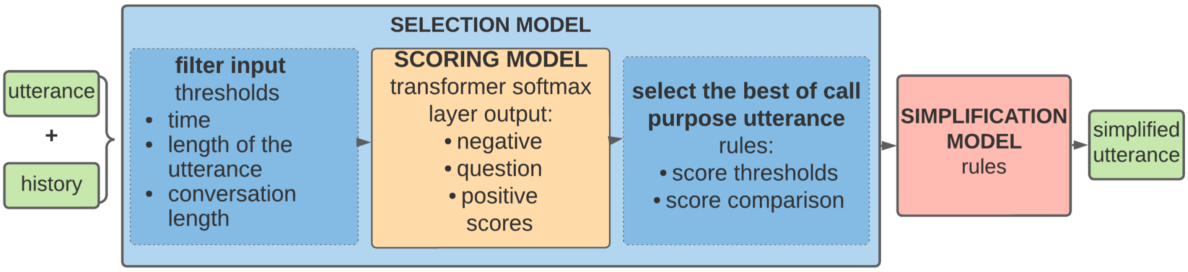


Figure 2: Purpose of Call Detection System Architecture

simplification with precision of 96%.

3.2 Model Development

The model development process consists of three main stages: data analysis and feature selection, designing a knowledge-engineered model informed by the insights from the data, and bootstrapping a transformer based model from the rule-based system. In this section, we discuss these stages.

3.2.1 Data Research and Feature Selection

We manually analyzed a sample of 2000 call transcripts across several dimensions, which are outlined below.

1. **Inbound vs. outbound calls**:
   In *outbound* call center calls (40.8%), the call initiator and the side that utters the Purpose of Call is usually an agent; in *inbound* calls (59.2%), it is a customer, with some exceptions. 55.3% of all Purpose of Call statements are uttered by the customer.

2. **Place**: the Purpose of Call is uttered within the first 40 seconds in a majority of calls (73.8%), in the middle in a minority of calls (10.7%), and towards the end in a handful of (mainly short) calls. The *mean* time of occurrence is 29.9 seconds, *std*=19.1, *median*=25.5 seconds. The *maximum* time is 180 seconds.

3. **Speaker role**:
   The *call initiator* utters the Purpose of Call in the vast majority of cases in the form of a *statement*; in returned or scheduled calls, the Purpose of Call can be uttered by a *call recipient* in the form of a *guess, assumption or inquiry*.

4. **Domain**: There are three main types of call center calls:
   Sales calls: commonly characterized by the Purpose of Call *not* being stated explicitly in one utterance, but gradually being revealed during the course of the call. Agents often spend a longer time building rapport, so utter the purpose later in the call compared to support calls. 74% of Purpose of Call statements still occur within the first 40 seconds. Outbound calls are prevalent. 48% of Purpose of Call statements are uttered by customers.
   Support calls: inbound calls are prevalent, the Purpose of Call is introduced early in the conversation. In fact, 56% of Purpose of Call statements are in the very first utterance, accompanied by a *greeting*, and 63% of the statements occur within the first 50 seconds. 62% of Purpose of Call statements are uttered by customers.
   General business calls: may include support and sales calls as well as other communications, both formal and informal. The Purpose of Call is often implied (e.g. a conversation between colleagues, transfers from a chat to a call, with the purpose of the conversation being known by both parties). 84% occur within the first 45 seconds, and 59% are uttered by customers.

5. **Length distribution**: Purpose of Call utterances range in length from 4 to 224 tokens, with the *mean*=45.5, *std*=29.9, *median*=37, and 75% being under 59 tokens.

6. **Language patterns**: We identified several language markers associated with Purpose of Call statements in Table 1.

Approximately 7% of calls in this sample do not contain an explicit Purpose of Call statement. Instead, the participants in the call appear to already have the context necessary to understand the call purpose.

Pattern	%	Description	Example
call_purpose_phrases	32.7	explicit declarations of the Purpose of Call typically signposted with lexical cues containing *purpose* and *call* and their synonyms	*The* ***reason for my call*** *is I moved to a new address, so I need to change it on my profile.*
desire_phrases	31.7	expressions of volition, desire or need	*Hi, I* ***need*** *a refund for my order.*
question_response	15.8	responses to an agent's prompt	*-* ***How can I help you?*** */ - I received a message that my order has been delayed.*
greetings	9.1	long statements of at least 30 word tokens that follow a *greeting* and occur within the first 6 utterances in the conversation	*Hey, this is Christine. There is a police report, it was next to you guys why you heard it <...>*
problem_phrases	4.4	express problems and concerns	*I'm having an issue with the delivery.*
update	5.8	updates and announcements	***I have an update*** *on your passport status.*
continuation	0.4	questions preceded by a signpost in the same utterance or a subsequent one from the same speaker	*Hi, I'm calling because* ***I have a question****. / Do you accept new patients?*

Table 1: Language patterns in Purpose of Call statements

3.2.2 Knowledge-Engineered Model

As outlined in Section 1, collecting labeled data for Purpose of Call extraction is a challenging task. Therefore, to obtain a representative sample of training data, we first implemented a knowledge-based model that takes into account the following parameters: *utterance length* in tokens, the *order* of an utterance in the conversation, the *history* including several preceding utterances, and the presence or absence of *language patterns* summarized in Table 1 and implemented using regular expressions syntax (see Appendix A for an example). In total, stemming from the analysis in Section 3.2.1, 8 rules (56 regex patterns) to detect call purpose candidates and 6 rules (55 regex patterns) to filter out negative statements were developed. The model reached a precision of 90.8% and hit rate of 77% on average across three domains (see Table 4).

Further, we conducted error analysis by manually labelling the output of the production system on a random sample of 1000 calls. After human review, we determined that 3% of calls did not contain an identifiable Purpose of Call and could be considered true negatives, while 20% were false negatives. 40% of these false negatives can be attributed to ASR errors. 27.6% of false negatives include cases with the Purpose of Call being known prior to the conversation (e.g. from shared knowledge, logged information, or in return calls) and therefore not considered by the model, 9.1% correspond to specific industries (e.g. transportation) underrepresented in the data used in the analysis, and 44.7% were caused by the limitations of the rules (note that these groups of false negatives intersect, hence the percentages do not add up to 100%). False positives were mainly related to the lack of morphological flexibility in the rules and speech dysfluencies. In 6.2% of calls, several utterances were legitimate Purpose of Call statements and the one selected by the model was not the best one. These findings motivated the need for a transformer-based model that was more forgiving of ASR noise, had better generalization power, and was more responsive to changes in the data.

3.2.3 Transformer-Based Model

Training Data Collection. Since the knowledge-engineered model achieved high precision, we could rely on its output to train a deep learning model. The dataset consists of English language utterances obtained from business calls in a variety of industries, with accompanying metadata such as timestamps for each token, call side, and call id. See Appendix C and 3.2.1 for detailed statistics. We randomly sampled one million utterances between February 6, 2020 and February 22, 2021, allowing only those that meet the requirements for a Purpose of Call candidate in 3.1. The utterances were divided into two sets: (1) those from the calls with a Purpose of Call hit (likely to be a true positive), (2) utterances from calls with no hit (may contain false negatives). With a series of patterns, we filtered out utterances that are likely to be false positives based on error analysis in 3.2.2. Further, we sampled several datasets of 180K utterances with varying label and language pattern distributions in order to experimentally find the best configuration (see Appendix C). A train, development, and validation split of 80/10/10% of data was used in each experiment. In addition, we created a golden dataset of 909 manually labeled calls, with the utterances organized chronologically within the call and limited to up to 30 utterances per call. This sample comprises 13215 utterances.

Training Details. We employ the DistilBERT (Sanh et al., 2019) model, trained for classification with multimodal features. We combine text features with numerical and binary features, utterance *start time* and *call side* respectively, which have proven to be useful in the knowledge-engineered model, and pass on a gated summation of the transformer output with these features to the classification layer, following the approach in (Gu and Budhkar, 2021). This configuration outperformed other base models [2] [3], combinations of multimodal features, and combining mechanisms outlined in (Gu and Budhkar, 2021). The model architecture is shown in Figure 3. We implement a data-driven iterative fine-tuning process with extensive error analysis and data resampling. See Appendix D for details.

3.3 Model Deployment

Since the model was to operate in a near-realtime environment as a call is ongoing, optimising for inference time was a dominant consideration during model design. The model would perform inference on one CPU core. The model would need to accommodate the time taken to transcribe voice to text and properly format and punctuate the transcription, many of these tasks being accomplished by other deep learning models.

Optimizations include: (i) Having the Selection model that uses input features to filter utterances, thereby reducing the number of utterances that were attended to by the transformer model. These features include utterance count number and utterance start time, both of which should be below a threshold determined by experimenting with different parameter values in the knowledge-engineered model. (ii) Incorporating numerical and binary features into the deep learning model - adding signals beyond lexical features allowed us to use a lower capacity BERT variant with faster inference time. (iii) Capping input length to an empirically derived ceiling further reduced memory consumption and inference time.

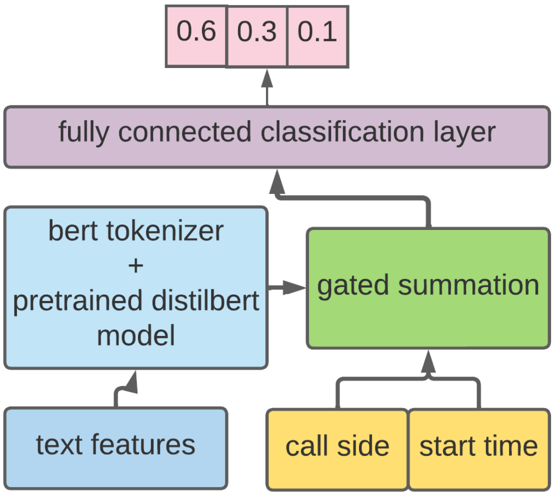


Figure 3: Multimodal Transformer based scoring model

The system was deployed in containers[4] with 1 CPU and maximum 1GB memory per instance. The average inference time at the 95th percentile is 0.51 seconds, which meets the requirements of our production system for near-realtime deployments to complete inference in under 3 seconds.

4 Evaluation

Table 2 shows full results of the comparative evaluation of the knowledge-engineered and the hybrid models on business calls in three domains.

Qualitative analysis was conducted on the gold set: real life user data of 600 samples, and 200 calls with missed hits. False positives mainly correspond to signposting language without mention of the actual Purpose of Call, and indicate the model's overreliance on lexical features. The model was found to be accurate in assigning utterances to classes, but not always sensitive to the difference between a

[2] https://huggingface.co/microsoft/DialoGPT-small

[3] https://huggingface.co/DeepPavlov/bert-base-cased-conversational

[4] https://cloud.google.com/kubernetes-engine

Domain	Model	Precision	Hit Rate	F1
Support	rules	93.5	80.0	86.2
	hybrid	91.0	90.4	90.7
General	rules	90.0	74.2	81.3
	hybrid	89.0	85.6	87.3
Sales	rules	88.5	78.7	83.5
	hybrid	87.0	88.9	87.9
Avg	rules	90.6	77.6	83.6
	hybrid	89.6	88.3	88.6

Table 2: Comparative evaluation of knowledge-engineered (here *rules*) and hybrid models for Purpose of Call detection.

valid and *the best* Purpose of Call. This can be addressed by introducing more contrastive examples in training data. Missed hits include cases initially excluded from the sample such as Purpose of Call stated across several utterances, and multiple Purpose of Call statements of equal importance. A synthesis of several utterances instead of selecting only one of them might be useful in such cases.

5 Conclusion

This paper discusses the development and deployment of a hybrid system to detect a Purpose of Call statement in business call transcripts for the English language in near-realtime settings. We introduce the concept of the Purpose of Call, provide in-depth analysis of real life data, and discuss overcoming the absence of available training data by bootstrapping from a knowledge-engineered model to a deep learning one. Both the knowledge-engineered and hybrid models demonstrate high precision and hit rate, with the hybrid model showing better performance while maintaining computational efficiency.

6 Ethics Statement

Data. The conversational data is presented in the form of individual utterances with sensitive data such as personal identifiable information removed. No crowdsourced annotation has been conducted, and access to the data was available only to a small number of in-house Scientists.

Use. The Purpose of Call feature is used by call center managers to identify areas to coach sales and support agents. It is recommended to not use this feature for automated evaluation of agent performance. Incorrect Purpose of Call prediction may provide unsatisfactory user experience for the managers as they sample calls but does not present any risk of negative impact for the agents.

Licensing. We follow the licensing requirements accordingly while using external tools such as HuggingFace [5] and Multimodal-Toolkit (Gu and Budhkar, 2021) libraries.

References

Frédérik Cailliau and Ariane Cavet. 2013. Mining automatic speech transcripts for the retrieval of problematic calls. In *CICLing*.

Chloé Clavel, Gilles Adda, Frédérik Cailliau, Martine Garnier-Rizet, Ariane Cavet, Géraldine Chapuis, Sandrine Courcinous, Charlotte Danesi, Anne-Laure Daquo, Myrtille Deldossi, Sylvie Guillemin-Lanne, Marjorie Seizou, and Philippe Suignard. 2013. Spontaneous speech and opinion detection: mining call-centre transcripts. *Language Resources and Evaluation*, 47:1089–1125.

Camille Dutrey, Chloé Clavel, Sophie Rosset, Ioana Vasilescu, and Martine Adda-Decker. 2014. A CRF-Based Approach to Automatic Disfluency Detection in a French Call-Centre Corpus. In *15th Annual Conference of the International Speech Communication Association (Interspeech'14)*, pages 2897–2901, Singapour, Singapore. International Speech Communication Association (ISCA).

Ken Gu and Akshay Budhkar. 2021. A package for learning on tabular and text data with transformers. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 69–73, Mexico City, Mexico. Association for Computational Linguistics.

Judith Hall, Phil Verghis, William Stockton, and Jin Goh. 2014. It takes just 120 seconds: Predicting satisfaction in technical support calls. *Psychology and Marketing*, 31.

Anita Pomerantz and B.J. Fehr. 2011. Conversation analysis: An approach to the analysis of social interaction. *Discourse studies: A multidisciplinary introduction*, pages 165–190.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Emanuel Schegloff and Harvey Sacks. 1973. Opening up closings. *Semiotica*, 8:289–327.

Hironori Takeuchi, L Venkata Subramaniam, Tetsuya Nasukawa, and Shourya Roy. 2007. Automatic identification of important segments and expressions for

[5] https://huggingface.co/

mining of business-oriented conversations at contact centers. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 458–467, Prague, Czech Republic. Association for Computational Linguistics.

Carmen Taleghani-Nikazm. 2002. A conversation analytical study of telephone conversation openings between native and nonnative speakers. *Journal of Pragmatics*, 34:1807–1832.

Piotr Żelasko, Jan Mizgajski, Mikołaj Morzy, Adrian Szymczak, Piotr Szymański, Łukasz Augustyniak, and Yishay Carmiel. 2019. Towards better understanding of spontaneous conversations: Overcoming automatic speech recognition errors with intent recognition.

A Appendix: Example Rule in a Knowledge-Engineered Model

An utterance is a Purpose of Call if:

- It contains signposting phrases expressing a problem such as *I'm having a problem, There is an issue, I'm having a hard time, I'm trying ... and it's not working*,
- It occurs within the first 10 utterances
- It is at least 12 tokens long

Example: *I got a really big problem here. When I log in, it asks for some pin, and I really, I can't use it. So there's obviously an issue here and can you help me with it?*.

B Appendix: Example Selection model heuristics

Combine the *positive* score for an utterance with the maximum *question* score of the two preceding utterances in another call side. If it passes a threshold and is the biggest score so far, this utterance is a Purpose of Call.

C Appendix: Data Statistics

Total number of calls: 86310
Total number of utterances: 180 000
Industry distribution: see Table 3.

Label distribution: A key factor in training the model was determining the right distribution of labels and language patterns. The classes in our problem are naturally imbalanced: since only one utterance per call is a valid Purpose of Call, the vast majority of utterances are of the *negative* class. In a random sample, only 4.7% utterances are *positive* hits, and only 1.9% are *questions*. If the data is sampled randomly, the model is likely to overfit to the *negative* class. Sampling uniformly may reduce the number of complex instances in favor of the ones easier for the model to learn. A set of experiments were conducted to determine the distribution of classes with the goal of optimizing accuracy of the Purpose of Call class predictions. We determined the optimal distribution of classes as follows: 42.5% *positive*, 42.5% *negative*, 15% *question* utterances (corresponds to the share of this pattern in real data). All utterances came from calls with a positive hit, which minimized the chance of false negatives in the training data.

Industry	%
Technology	25.1
IT, Consulting	15.5
Professional, Business Support Services	14.1
Travel	11.4
Health and Wellness	5.6
Real Estate	5.1

Table 3: Industry distribution in training data: top 6 types

Language patterns distribution: From the error analysis and experiments, we determined the optimal distribution of language patterns within the *positive* class:

- 30% *call_purpose_phrases*
- 30% *desire_phrases*
- 20% *problem_phrases*
- 20% *other patterns*

Other aspects of the data are the same as described in 3.2.1.

D Appendix: Training details

Parameters: The pretrained *distilbert-base-cased* model we use has 6 layers, 768 hidden units, 12 attention heads and 65M parameters and is available through Multimodal-Toolkit (Gu and Budhkar, 2021). We run all fine-tuning experiments on a Google Cloud VM n1-standard-8 instance with 496GB disk size and 1 NVIDIA Tesla K80 GPU. The maximum time for a single experiment was 8 GPU hours. We truncate text input to a maximum 150 tokens since most relevant statements fall into this category. We set the train and eval batch size to

32 and 64 respectively, and use AdamW optimizer with default parameters. We fine-tune the model for 4 epochs with a learning rate of 5e-05, weight decay of 0.01 and 500 warm up steps. The hyperparameters were obtained from experiments using an in-house tuning tool implementing grid search algorithm. For fine-tuning on small subsets (4K) of data collected through error analysis, we repeat the training process for 12 epochs and a learning rate of 9e-05.

Relevant features: Besides the text features, we experimented with two extra features that have proven to be useful in the knowledge-engineered model: the *start time* of the utterance and the *call side*. We also experimented with several mechanisms to combine the numerical and categorical features with textual data using Multimodal-Toolkit (Gu and Budhkar, 2021). The results are presented in Table 4.

Feature	**P**	**HR**	**F1**	**PP**
text only	0.891	0.891	0.891	0.894
text + start time	0.948	0.948	0.948	0.944
text + call side	0.948	0.948	0.948	0.946
all	0.949	0.949	0.949	0.957

Table 4: Comparing model performance using tabular features *start time* and *call side*. P-Precision, HR-Hit rate, PP - precision in *positive* class. The results are reported for a single run using concatenation to combine features.