Alex: Bootstrapping a Spoken Dialogue System for a New Domain by Real Users*

Ondřej Dušek, Ondřej Plátek, Lukáš Žilka, and **Filip Jurčíček**
Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, CZ-11800 Prague, Czech Republic
{odusek,oplatek,zilka,jurcicek}@ufal.mff.cuni.cz

Abstract

When deploying a spoken dialogue system in a new domain, one faces a situation where little to no data is available to train domain-specific statistical models. We describe our experience with bootstrapping a dialogue system for public transit and weather information in real-word deployment under public use. We proceeded incrementally, starting from a minimal system put on a toll-free telephone number to collect speech data. We were able to incorporate statistical modules trained on collected data – in-domain speech recognition language models and spoken language understanding – while simultaneously extending the domain, making use of automatically generated semantic annotation. Our approach shows that a successful system can be built with minimal effort and no in-domain data at hand.

1 Introduction

The Alex Public Transit Information System is an experimental Czech spoken dialogue system providing information about all kinds of public transit in the Czech Republic, publicly available at a toll-free 800 telephone number.[1] It was launched for public use as soon as a first minimal working version was developed, using no in-domain speech data. We chose an incremental approach to system development in order to collect call data and use them to bootstrap statistical modules. Nearly a year after launch, we have collected over 1,300 calls from the general public, which enabled us to train and deploy an in-domain language model for Automatic Speech Recognition (ASR) and a statistical Spoken Language Understanding (SLU) module. The domain supported by the system has extended from transit information in one city to ca. 5,000 towns and cities in the whole country, plus weather and time information. This shows that a even a very basic system is useful in collecting in-domain data and that the incremental approach is viable.

Spoken dialogue systems have been a topic of research for the past several decades, and many experimental systems were developed and tested with users (Walker et al., 2001; Gašić et al., 2013; Janarthanam et al., 2013). However, few experimental systems became available to general public use. Let's Go (Raux et al., 2005; Raux et al., 2006) is a notable example in the public transportation domain. Using interaction with users from the public to bootstrap data-driven methods and improve the system is also not a common practice. Both Let's Go and the GOOG-411 business finder system (Bacchiani et al., 2008) collected speech data, but applied data-driven methods only to improve statistical ASR. We use the call data for statistical SLU as well and plan to further introduce statistical modules for dialogue management and natural language generation.

Our spoken dialogue system framework is freely available on GitHub[2] and designed for easy adaptation to new domains and languages. An English version of our system is in preparation.

We first present the overall structure of the Alex SDS framework and then describe the minimal system that has been put to public use, as well as our incremental extensions. Finally, we provide an evaluation of our system based on the recorded calls.

*This work was funded by the Ministry of Education, Youth and Sports of the Czech Republic under the grant agreement LK11221 and core research funding, SVV project 260 104, and grants GAUK 2058214 and 2076214 of Charles University in Prague. It used language resources stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).

[1] Call 800-899-998 from the Czech Republic.

[2] http://github.com/UFAL-DSG/alex

Proceedings of the SIGDIAL 2014 Conference, pages 79–83,
Philadelphia, U.S.A., 18-20 June 2014.

2 Overall Alex SDS System Structure

The basic architecture of Alex is modular and consists of the traditional SDS components: automatic speech recognizer (ASR), spoken language understanding (SLU), dialogue manager (DM), natural language generator (NLG), and a text-to-speech (TTS) module.

We designed the system to allow for easy replacement of the individual components: There is a defined interface for each of them. As the interfaces are domain-independent, changing the domain is facilitated as well by this approach.

3 Baseline Transit Information System

We decided to create a minimal working system that would not require any in-domain data and open it to general public to collect call data as soon as possible. We believe that this is a viable alternative to Wizard-of-Oz experiments (Rieser and Lemon, 2008), allowing for incremental development and producing data that correspond to real usage scenarios (see Section 4).

3.1 Baseline Implementation of the Components

Having no in-domain data available, we resorted to very basic implementations using hand-written rules or external services:

- ASR used a neural network based voice activity detector trained on small out-of-domain data. Recordings classified as speech were fed to the the web-based Google ASR service.
- SLU was handcrafted for our domain using simple keyword-spotting rules.
- In DM, the dialogue tracker held only one value per dialogue slot, and the dialogue policy was handcrafted for the basic tasks in our domain.
- NLG is a simple template-based module.
- We use a web-based Czech TTS service provided to us by SpeechTech.[3]

3.2 Baseline Domain

At baseline, our domain only consisted of a very basic public transport information for the city of Prague. Our ontology contained ca. 2,500 public transit stops. The system was able to present the next connection between two stops requested by the user, repeat the information, or return several following connections. Connection search was based on Google Directions API.[4]

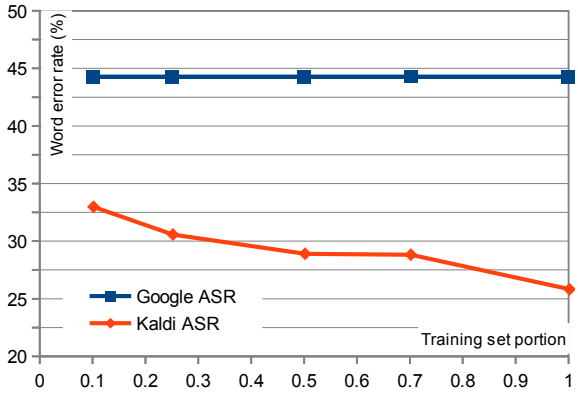

Figure 1: ASR word error rate depending on the size of in-domain language model training data

The full training set amounts to 9,495 utterances (30,126 tokens). The test set contains 1,187 utterances (4,392 tokens).

4 Collecting Data and Extending the System in Real Usage

We launched our system at a public toll-free 800 number and advertised the service at our university, among friends, and via Facebook. We also cooperate with the Czech Blind United association,[5] promoting our system among its members and receiving comments about its use. We advertised our extensions and improvements using the same channels.

We record and collect all calls to the system, including our own testing calls, to obtain training data and build statistical models into our system.

4.1 Speech Recognition: Building In-Domain Models

The Google on-line ASR service, while reaching state-of-the-art performance in some tasks (Morbini et al., 2013), showed very high word error rate in our specific domain (see Figure 1). We replaced it with the Kaldi ASR engine (Povey et al., 2011) trained on general-domain Czech acoustic data (Korvas et al., 2014) with an in-domain class-based language model built using collected call data and lists of all available cities and stops.

We describe our modifications to Kaldi for on-line decoding in Plátek and Jurčíček (2014). A performance comparison of Google ASR with

[3] http://www.speechtech.cz/

[4] https://developers.google.com/maps/documentation/directions/

[5] http://www.sons.cz

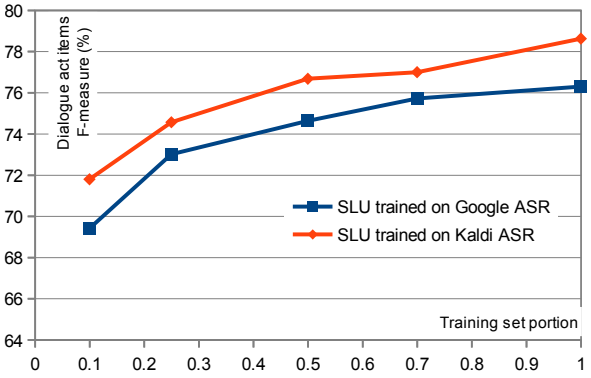

Figure 2: SLU performance (F-measure on dialogue act items) depending on training data size

The same data sets as in Figure 1 are used, with semantic annotations from handcrafted SLU running on manual transcriptions.

Kaldi trained on our data is shown in Figure 1. One can see that the in-domain language model brings a substantial improvement, even with very small data sizes.

4.2 Spoken Language Understanding

To increase system robustness, we built a statistical SLU based on a set of logistic regression classifiers and word n-gram features (Jurčíček et al., 2014). We train it on the output of our handcrafted SLU applied to manual transcriptions. We chose this approach over obtaining manual semantic annotation due to two main reasons:

1. Obtaining semantic annotation for Czech data is relatively slow and complicated; using crowdsourcing is not a possibility due to lack of speakers of Czech on the platforms.
2. As we intended to gradually extend our domain, semantic annotation changed over time as well.

This approach still allows the statistical SLU to improve on a handcrafted one by compensating for errors made by the ASR. Figure 2 shows that the performance of the statistical SLU module increases with more training data and with the in-domain ASR models.

4.3 Dialogue Manager

We have replaced the initial simplistic dialogue state tracker (see Section 3.1) by the probabilistic discriminative tracker of Žilka et al. (2013), which achieves near state-of-the-art performance while remaining completely parameter-free. This property allowed us to employ the tracker without any training data; our gradual domain extensions also required no further adjustments.

The dialogue policy is handcrafted, though it takes advantage of uncertainty estimated by the belief tracker. Its main logic is similar to that of Jurčíček et al. (2012). First, it implements a set of domain-independent actions, such as:

- dialogue opening, closing, and restart,
- implicit confirmation of changed slots with high probability of the most probable value,
- explicit confirmation for slots with a lower probability of the most probable value,
- a choice among two similarly probable values.

Second, domain-specific actions are implemented for the domain(s) described in Section 4.4.

4.4 Extending the Domain

We have expanded our public transit information domain with the following tasks:

- The user may specify departure or arrival time in absolute or relative terms (“in ten minutes”, “tomorrow morning”, “at 6 pm.”, “at 8:35” etc.).
- The user may request more details about the connection: number of transfers, journey duration, departure and arrival time.
- The user may travel not only among public transport stops within one city, but also among multiple cities or towns.

The expansion to multiple cities has lead to an ontology improvement: The system is able to find the corresponding city in the database based on a stop name, and can use a default stop for a given city. We initially supported three Czech major cities covered by the Google Directions service, then extended the coverage to the whole country (ca. 44,000 stops in 5,000 cities and towns) using Czech national public transport database provided by CHAPS.[6]

We now also include weather information for all Czech cities in the system. The user may ask for weather at the given time or on the whole day. We use OpenWeatherMap as our data source.[7]

Furthermore, the user may ask about the current time at any point in the dialogue.

5 System Evaluation from Recorded Calls

We have used the recorded call data for an evaluation of our system. Figure 3 presents the num-

[6] http://www.idos.cz
[7] http://openweathermap.org/

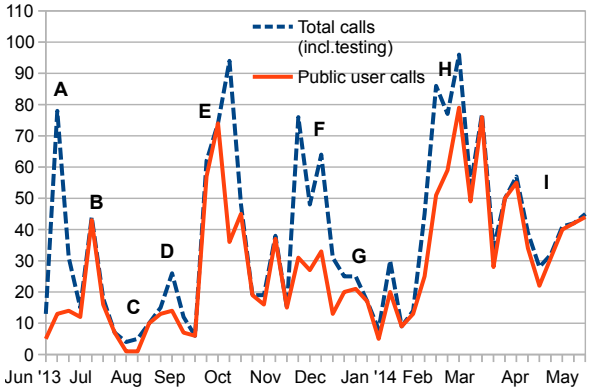

Figure 3: Number of calls per week

The dashed line shows all recorded calls, including those made by the authors. The full line shows calls from the public only.
Spikes: *A* – initial testing, *B* – first advertising, *C* – system partially offline due to a bug, *D* – testing statistical SLU module, *E* – larger advertising with Czech Blind United, *F* – testing domain enhancements, *G* – no advertising and limited system performance, *H* – deploying Kaldi ASR and nationwide coverage, *I* – no further advertising.

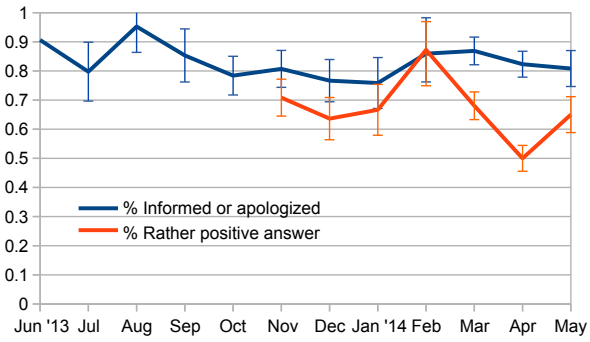

Figure 4: System success rates by month

Percentage of calls where the system provided information (or apology for not having one) and percentage of rather positive responses to the final question, both shown with standard error bars.

ber of calls to our system per week and reflects the testing and advertising phases, as well as some of our extensions and improvements described in Section 4. A steeper usage increase is visible in recent weeks after the introduction of Kaldi ASR engine and nationwide coverage (see Sections 4.1 and 4.4). The number of calls and unique users (caller phone numbers) grows steadily; so far, more than 300 users from the public have made over 1,300 calls to the system (cf. Figure 5 and Table 1 in the appendix).[8]

Figure 4 (and Table 1 in the appendix) give a detailed view of the success of our system. Information is provided in the vast majority of calls. Upon manual inspection of call transcripts, we discovered that about half of the cases where no information is provided can be attributed to the system failing to react properly; the rest is off-topic calls or users hanging up too early.

We have also introduced a "final question" as an additional success metric. After the user says good-bye, the system asks them if they received the information they were looking for. By looking at the transcriptions of responses to this question, we recognize a majority of them as rather positive ("Yes", "Nearly" etc.); the proportion of positive reactions seems to remain stable. However, the final question is not an accurate measure as most users seem to hang up directly after receiving information from the system.

6 Conclusions and Further Work

We use an iterative approach to build a complex dialogue system within the public transit information domain. The system is publicly available on a toll-free phone number. Our extensible dialogue system framework as well as the system implementation for our domain can be downloaded from GitHub under the Apache 2.0 license.

We have shown that even very limited working version can be used to collect calls from the public, gathering training data for statistical system components. Our experiments with the Kaldi speech recognizer show that already a small amount of in-domain data for the language model brings a substantial improvement. Generating automatic semantic annotation from recording transcripts allows us to maintain a statistical spoken language understanding unit with changing domain and growing data.

The analysis of our call logs shows that our system is able to provide information in the vast majority of cases. Success rating provided by the users themselves is mostly positive, yet the conclusiveness of this metric is limited as users tend to hang up directly after receiving information.

In future, we plan to add an English version of the system and further expand the domain, allowing more specific connection options. As we gather more training data, we plan to introduce statistical modules into the remaining system components.

[8] We only count calls with at least one valid user utterance, disregarding calls where users hang up immediately.

A System Evaluation Data

In the following, we include additional data from call logs evaluation presented in Section 5.

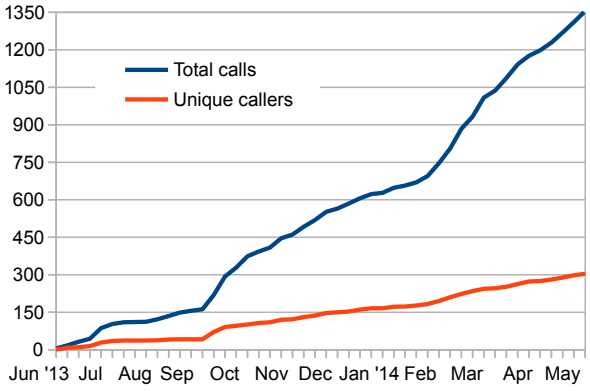

Figure 5: Cumulative number of calls and unique callers from the public by weeks

The growth rates of the number of unique users and the total number of calls both correspond to the testing and advertising periods shown in Figure 3.

Total calls	1,359
Unique users (caller phone numbers)	304
System informed (or apologized)	1,124
System informed about directions	990
System informed about weather	88
System informed about current time	41
Apologized for not having information	223
System asked the final question	229
Final question answered by the user	199
Rather positive user's answer	146
Rather negative user's answer	23

Table 1: Detailed call statistics

Total absolute numbers of calls from general public users over the period of nearly one year are shown.

References

M. Bacchiani, F. Beaufays, J. Schalkwyk, M. Schuster, and B. Strope. 2008. Deploying GOOG-411: early lessons in data, measurement, and testing. In *Proceedings of ICASSP*, page 5260–5263. IEEE.

M. Gašić, C. Breslin, M. Henderson, D. Kim, M. Szummer, B. Thomson, P. Tsiakoulis, and S. Young. 2013. On-line policy optimisation of bayesian spoken dialogue systems via human interaction. In *Proceedings of ICASSP*, page 8367–8371. IEEE.

S. Janarthanam, O. Lemon, P. Bartie, T. Dalmas, A. Dickinson, X. Liu, W. Mackaness, and B. Webber. 2013. Evaluating a city exploration dialogue system combining question-answering and pedestrian navigation. In *Proceedings of ACL*.

F. Jurčíček, B. Thomson, and S. Young. 2012. Reinforcement learning for parameter estimation in statistical spoken dialogue systems. *Computer Speech & Language*, 26(3):168–192.

F. Jurčíček, O. Dušek, and O. Plátek. 2014. A factored discriminative spoken language understanding for spoken dialogue systems. In *Proceedings of TSD*. To appear.

M. Korvas, O. Plátek, O. Dušek, L. Žilka, and F. Jurčíček. 2014. Free English and Czech telephone speech corpus shared under the CC-BY-SA 3.0 license. In *Proceedings of LREC*, Reykjavík.

F. Morbini, K. Audhkhasi, K. Sagae, R. Artstein, D. Can, P. Georgiou, S. Narayanan, A. Leuski, and D. Traum. 2013. Which ASR should i choose for my dialogue system? In *Proceedings of SIGDIAL*, page 394–403.

O. Plátek and F. Jurčíček. 2014. Free on-line speech recogniser based on kaldi ASR toolkit producing word posterior lattices. In *Proceedings of SIGDIAL*.

D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al. 2011. The Kaldi speech recognition toolkit. In *Proceedings of ASRU*, page 1–4, Hawaii.

A. Raux, B. Langner, D. Bohus, Alan W. Black, and M. Eskenazi. 2005. Let's go public! taking a spoken dialog system to the real world. In *Proceedings of Interspeech*.

A. Raux, D. Bohus, B. Langner, Alan W. Black, and M. Eskenazi. 2006. Doing research on a deployed spoken dialogue system: one year of Let's Go! experience. In *Proceedings of Interspeech*.

V. Rieser and O. Lemon. 2008. Learning effective multimodal dialogue strategies from Wizard-of-Oz data: Bootstrapping and evaluation. In *Proceedings of ACL*, page 638–646.

M. A. Walker, R. Passonneau, and J. E. Boland. 2001. Quantitative and qualitative evaluation of DARPA communicator spoken dialogue systems. In *Proceedings of ACL*, page 515–522.

L. Žilka, D. Marek, M. Korvas, and F. Jurčíček. 2013. Comparison of bayesian discriminative and generative models for dialogue state tracking. In *Proceedings of SIGDIAL*, page 452–456, Metz, France.