A UNIVERSAL MODEL FOR FLEXIBLE ITEM SELECTION IN CONVERSATIONAL DIALOGS

Asli Celikyilmaz, Zhaleh Feizollahi, Dilek Hakkani-Tur, Ruhi Sarikaya

Microsoft

ABSTRACT

Human-computer interaction and statistical natural language understanding has changed with the addition of a visual display screen in modern mobile devices, as visual rendering is used to communicate the dialog system's response. On-screen item identification and resolution when interpreting the user utterances is one critical problem to achieve the natural and accurate human-machine communication. This problem, also called Flexible Item Selection (FIS), has been posed as a classification task to correctly identify intended on-screen item(s) from user utterances. This paper presents a universal FIS model that can be applied to dialog systems developed in different languages. We design a set of input features for the FIS model that makes it largely language-independent. We demonstrate that a single universal FIS model can be used in place of language specific FIS models with no loss in accuracy. We also show that such a model can generalize well to new unseen languages with minimal loss in accuracy on held out languages including English, French, Spanish, Italian, German, and Chinese. Eliminating the need for building and maintaining a separate FIS model for each new language, the universal FIS model helps scaling an existing dialogue system to new languages faster at a lower development cost.

Index Terms— on screen item selection, multi language and universal models, language expansion, spoken language understanding, spoken dialog systems, language independence.

1. INTRODUCTION

Spoken language understanding systems (e.g. SIRI, Cortana, Google Now) running on smart phones and consoles (e.g. Xbox) provide a natural user interface (NUI), enabling a more natural interaction for the user. As NUI's become mainstream, scaling such applications to different languages has created new and challenging problems in spoken dialog systems. This paper focuses on *building language independent Flexible Item Selection (FIS) for spoken dialog systems to identify and resolve on-screen item given a user query.*

We previously introduced a FIS model in [1]. The model is learned by associating the user utterances with items on the screen. Consider a sample dialog in Table 1 between a user and spoken dialog system in a '*location search*' domain that covers restaurants as a subset. After the system displays results on the screen in response to the first turn user query, the user may choose one or more of the on-screen items. Note that, there are multiple ways of referring to the same item[1].

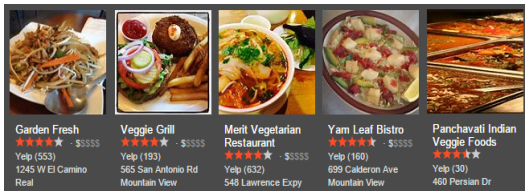
[System]: How may i help you today ?
[User]: *Show me veggie places near me.*
[System]: (Fetches the following restaurants)

[User]: ***"show menu for the Chinese one"*** or ***"directions to the nearest one"*** or ***"the one on el camino"*** or ***"Saturday hours for the first one"*** or *"is there a Starbucks nearby"*

Table 1. A sample multi-turn dialog. A list of second turn utterances (in bold) all referring to the first restaurant and a new search query (highlighted) are shown.

In a typical SDS, the spoken language understanding (SLU) engine maps user utterances into meaning representation by identifying user's intent and token level semantic slots via a semantic parser [2, 3, 4]. The dialog manager uses the SLU components to decide on the correct system action. For on-screen item selection SLU alone may not be sufficient. For instance, consider the dialog in Table 1. SLU engine can provide signals to the dialog model about the selected item, e.g., that "*chinese*" is a restaurant-cuisine or content, but may not be enough to indicate which restaurant the user is referring. FIS model provides additional information to the dialog manager by indicating whether there is a relation between the utterance and the item, representing each instance in the training dataset with relational features.

The input features to FIS models are language independent since the features are mostly derived in the semantic space, e.g. the existence of a slot tag but not the actual words tagged, or the position of the item and the positional reference in the utterance (e.g., "*first* one"). Provided that the set

[1]An item could be anything, e.g. restaurants, games, contact list, and organized in different lay-outs, such as a list or grid on the screen.

of domains handled by the dialogue systems are largely language independent, FIS models should generalize well across dialogue systems operating in different languages, including previously unseen languages. This, if true, is a useful property as it eases porting such systems to new languages in different domains.

In the next section, we will provide related work followed by details on FIS model and features. Later, we will introduce the universal FIS models in section 4. Experiments are discussed in section 5. Finally conclusions are drawn.

2. RELATED WORK

As important as understanding the user's flexible selection requests for modern NUI designs, relatively few studies have investigated their performance on the SDSs. Those that do focus on the impact of the input from multimodal interfaces such as gesture for understanding [5, 6, 7], touch for Automatic Speech Recognition (ASR) error correction [8], or cues from the screen [9, 10]. Most of these systems are engineered for a specific task, making it harder to generalize for other domains, languages or SDSs.

In [11] a reference resolution model is presented for a question-answering system on a mobile, multi-modal interface. Their system has several features to parse the posed question and keep history of the dialog to resolve co-reference issues. Their question-answering model uses gesture as features to resolve queries such as "*what's the name of that [pointing gesture] player*?", but they do not resolve locational referrals such as "*the middle one*" or "*the second harry potter movie*". Others such as in [12] resolve anaphoric ("*it*") or exophoric ("*this one*") types of expressions in user utterances to identify geometric objects. In this paper, we study several types of referring expressions to build a natural and flexible interaction for the user and design features that are language independent.

Several research focused on improving natural language understanding for multi-modal interfaces for spoken or text input. In [6] an intent prediction model enriched with gesture detector is presented to help disambiguate between different spoken user intents related to the interface. In [13, 14], the impact of eye gaze on spoken language understanding models are investigated. In [15] a situated in-car dialog model is presented to answer drivers' spoken queries about the surroundings (no display screen) by integrating multi-modal inputs of speech, geo-location and gaze. In [16] a smart selection is presented for point finger touch interface (personal tablet device), that tries to recover a user's intent from the selected text (a word or a phrase) on a screen showing a document page. In their work, no user utterances are used as input modality. We investigate automatic identification and resolution of referring expressions as a classification task introducing a wide range of syntactic, semantic and contextual features extracted from spoken dialog data on different languages. To the best of our knowledge, our FIS model is the first domain-independent and language-independent referring expression solution for handling on-screen items.

3. LANGUAGE SPECIFIC FLEXIBLE ITEM SELECTION

The language specific FIS (Flexible Item Selection) model [1] can detect if the user is referring to any item(s) on the screen, and later resolve referring expressions to identify which items are referred to and score each item. We have observed four types of referring expressions, however, real usage analysis of multi-model dialog systems have revealed that users mostly refer to the items on the screen using expressions relating to the implicit or explicit locational.

Explicit Referential: Explicit mentions of whole or portions of the item's title on the screen with no other descriptors, e.g.,"*show me the details of star wars six*" (referring to the item with title "Star wars: Episode VI - Return of the Jedi").

Implicit Referential : The user refers to the item using distinguishing features other than the title, such as the release or publishing date, actors, image content (describing the item image), genre, etc. "*how about the one with Kevin Spacey*".

Explicit Locational : The user refers to the item using the grid interface design, e.g., "*i want to watch the movie on the bottom right corner*".

Implicit Locational: References in relation to other items on the screen, e.g., *the second of Dan Brown's book*" (showing two of the Dan Brown's book on the same row).

3.1. Relational Feature Extraction

The FIS problem is casted as a classification problem to detect the relation between the utterance and the item, representing each instance in the training dataset as relational features. As summarized below, these features do not rely on any specific domain or genre, and thus are domain independent.

Similarity Features: Similarity features represent the lexical overlap between the utterance and the item's title (that is displayed on the user's screen) and are mainly aimed to resolve *explicit referring expressions*. Each utterance and item-title is represented as sequence of words. Since inflectional morphology may make a word appear in an utterance in a different form than what occurs in the official title, we use both the word form as it appears in the utterance and in the item title. For example, *burger* and *burgers*, or *woman* and *women* are considered as four distinct words and all included in the bag-of-words. Using this representation we calculate four different similarity measures:

Jaccard Similarity: A common feature that can represent the ratio of the intersection to the union of unigrams.

Orthographic Distance: Represents the similarity of two text and can be as simple as an edit distance (Levenshtein distance) between their graphemes. The Levenshtein distance

[17] counts the insertion, deletion and substitution operations that are required to transform an utterance into item's title.

Word Order: This feature represents the similarity between the order of words in two text. Sentences containing the same words but in different orders may result in different meanings. Extending Jaccard similarity, this feature looks for overlapping bi-grams in the utterance and item.

Word Vector: This feature is the cosine similarity between the utterance and the item-title that measures the cosine of the angle between their vectors.

Location Bearing Features: This feature set captures spatial cues in utterances and is mainly aimed to resolve *explicit locational* referring expressions. The SLU models aim at resolving a domain independent slot-tag, the *position-ref* for locational references. This tag indicates the tokens in utterances that are indicative of a position (e.g., "*first* one", or "*last* one"). A language dependent canonicalization engine first converts the *position-ref* slot value to a numerical value (e.g., "*first*" is converted to "1") and then checks against the position of the item on the screen. This simple conversion applies to screen designs where items are listed as a single row (as in Table 1) or as a list in single column. See [1] for a statistical approach designed to resolve locational cues in utterances for more complex screen designs (e.g., grid based).

SLU Features: The SLU (Spoken Language Understanding) features are used to resolve *implicit* and *explicit referring expressions*. The language specific SLU model is a multi-turn, multi-domain statistical model that consists of a set of semantic attributes from utterances: the domain, user's intent and semantic slots based on a pre-defined semantic schema. For each domain, the intents are determined using a multi-class SVM intent model. The best intent hypothesis is used as a categorical feature for the FIS model. Although FIS is not an intent detection model, the intent from SLU is an effective semantic feature in resolving referring expressions. Consider second turn utterance such as "*weather in seattle*", which is a '*find*' intent that is a new search or not related to any item on the screen. Finally entities (slots) are tagged using conditional random fields (CRF) [18]. The best slot hypothesis from the SLU slot model and the feature value is determined based on the full overlap of any recognized slot value with either the item's title or meta-information from knowledge source (address, year, etc.) Although the top intent may indicate whether the user is referring an item on the screen (does not resolve which item is being referred to), it is not a domain independent feature since in each domain the top intent might be different. On the other hand, the slot matching features are domain and language independent as the feature value is determined based on the partial or exact match between the recognized slot in utterance and the item text or its meta information.

Knowledge Feature: This binary feature is used to represent overlap between utterance and the meta information about the item and is mainly aimed at resolving *implicit referring expressions*. First, the meta information about the on-screen items using knowledge graph and other web sources are obtained. These correspond to the knowledge about classes (books, movies, ...) and their attributes (title, publisher, year-released, ...). For each semantic frame relevant knowledge, e.g. database hits, are fetched and appended. Then, this information is checked against the utterance for any overlaps. For instance, given an utterance "*how about the one with Kevin Spacey*", and the item-title "*House of Cards*", the knowledge graph attributes include year(*2013*), cast(*Kevin Spacey*), director(*James Foley*),... For the FIS models, we turn the knowledge graph feature 'on' since the actor attribute of that item is contained in the utterance. We also consider partial matches, e.g., last name of the actor attribute.

3.2. Learning the FIS Model

In this paper the FIS models are trained using GBDT (gradient boosted decision tree) [19] models, also known as MART (Multiple Additive Regression Trees). GBDT[2] is an efficient algorithm which learns an ensemble of trees and they are easy to interpret compared to other non-linear classifiers such SVM (support vector machines) [20] or NN (neural networks) [21]. In [1], GBDTs were shown to perform considerably better on language dependent FIS models than SVMs.

4. UNIVERSAL FLEXIBLE ITEM SELECTION

Our hypothesis for building a universal FIS model is that given a shared approach to SLU and back-end knowledge resources across all languages, a single universal language-independent FIS model used for dialog systems deployed in different languages should perform as accurately as the language-dependent FIS models. Furthermore, that such universal FIS model should generalize well to such a dialogue system operating in an unseen language.

To be a universal FIS model, the extracted features should not rely on language specific information in which the FIS model is deployed. Thus we carefully selected features so that none of the extracted features directly contain words or phrases from the user's utterance, e.g., no *n*-gram features. On the other hand, the SLU models trained per domain and language do use lexical features. Therefore, during feature extraction for universal FIS models, we deliberately decided to avoid using lexical features primarily to avoid the GBDT model from having to recompute the lower level lexical analysis already taken upon by the SLU models. We also eliminated intent-type specific features, (e.g., best intent hypothesis) and slot-type specific features, (e.g. best slot hypothesis). However, we kept the slot matching features, which are

[2]Treenet: http://www.salford-systems.com/products/ treenet is the implementation of the GBDT which is used in this paper.

Domain	Intents (I) & Slots
places	I: find-place, select-item(*first one*).. Slots: place-type, rating, nearby(*closest*)....
communication	I: send-email, send-text, Slots: email-subject, to-contact-name,message,
reminder	I: change-reminder-text, set-alarm, Slots: reminder-text, change-to-time, location,...

Table 2. A sample of intents and semantic slot tags of utterance segments per domain. Examples for some slots values are presented in parenthesis as *italicized*.

binary features indicating the presence or absence of a particular entity (slot) value (captured by the SLU slot models) in the item's title as well as the item's meta information - hence a domain independent relational feature.

The similarity features are not only domain but also language independent as they are solely based on the lexical similarity of words and word orders between the utterance and item. In addition, location bearing features also match the location bearing phrase in the utterance to the position of the item on the screen (as user sees). Although the position-ref slot value is SLU specific, which in turn is not language specific, similar to slot matching features, the position matching against screen location is essentially language independent.

5. EXPERIMENTS

5.1. Datasets

The internal corpora used for training and testing consists mostly of logs of spoken utterances or typed input collected from real users of Cortana − Microsoft's personal digital assistant. This is mixed with a much smaller fraction of manually engineered or crowd sourced data. The log data is segmented into sessions based primarily on when users closed the Cortana application.

Collection: Our experimentation data is collected for six different languages. Since our initial FIS model investigations were on American English (en-us), the en-us data is twice as much as the rest of the languages. Around 100K training utterances were collected for Spanish (es-es), and Chinese Mandarin (zh-cn) languages and around fifty-thousand training utterances were collected for the French (fr-fr), German (de-de), and Italian (it-it) languages. The corpora for all of the languages span three distinct domains (places, communications, reminder) with multiple intents per domain as shown in Table 2. Around 20K of these utterances are held out for testing purposes.

Six dialogue systems were set up, one corresponding to each of the language pairs for which user data was collected. Each of these dialogue systems has a different, language specific SLU models and language specific knowledge sources. To run the experiments in this paper, the transcribed utterances and typed input text were processed to match the expected form of the 1-best output of the ASR and then fed into the SLU component.

The training corpora were run through their corresponding dialogue system in each language just until the user utters the subsequent turn (turn-#$>$1 indicates utterances at dialog turns other than the first turn). Note that since no items are displayed on the screen when the user starts the conversation, the FIS model is only triggered on the follow-up turns. Once the system returns results and the user utters the second turn, the feature extraction stage for FIS models starts. Features are collected and stored from the set of query-item pairs, one set for each displayed item (see Table-1). The result is a set of training examples with input features required by the FIS models which are associated with selection or no-selection labels as supervisory signals. A separate test corpora is also collected for each language and processed in the same way. The test corpora is held out from SLU model training as well as FIS model training.

FIS Data Annotation: We use Microsoft's crowd-sourcing services to annotate the training utterance-item pairs. Particularly, the labelers are shown a screen shot showing the search results after a search query is (turn-# $>$1) issued. Since we are building a relational model between utterances and each item on the screen, we ask the annotators to label each utterance-item as '0' or '1' indicating if the utterance is referring to that item or not. '1' means the item is the intended one. '0' indicates the item is not intended one or the utterance is not referring to any item on the screen, e.g., new search query. For example, given the screen shot and the query "*the one on el camino*" from Table-1, the labeler is asked to tag the relation between utterance and item-1 as '1' and '0' for all the rest of the items.

5.2. Experiment Setup

We run three experiments to investigate the universal FIS models as follows:

Experiment-1: Universal Model Performance: In this experiment, we train a single, universal FIS model on the entire training corpora from all six languages. This model is then tested in each of the six dialogue systems using the language specific test set for that dialogue system. The universal FIS model is compared to those of language specific FIS models, each of which are trained solely on the language corpus that matches that dialogue system's language.

Experiment-2: Scalability of Universal Models to Unseen Languages: In the second experiment, we tested the scalability of the universal FIS models, specifically by testing the FIS models on unseen languages. One language's training data is completely held-out and a 'universal' FIS model trained using the data from the the remaining languages. This was repeated for all the languages. As a comparison each of the language specific FIS models trained in the previous experiment were also tested against non-matching languages in order to test the assumption that a universal FIS model would generalize bet-

Language language	Specific Accuracy	Universal FIS FIS Accuracy	Δ in Accuracy
en-us	94.74%	96.08%	**+1.34%**
fr-fr	93.81%	90.33%	-3.48%
es-es	92.45%	92.84%	**+0.39%**
it-it	95.39%	92.70%	-2.69%
de-de	95.37%	92.35%	-3.02%
zh-cn	94.64%	98.23%	**+3.59%**

Table 3. Accuracy in selection models of Universal FIS trained on all languages versus FIS models trained on one specific language. Δ in accuracy % is the absolute difference between the accuracy achieved by the universal FIS models (third column) and the language dependent FIS models (second column). The absolute gain by the universal models are bolded.

Held-Out Language	Universal FIS Accuracy (w/ held-out)	Universal FIS Accuracy (w/o held-out)	Δ in Accuracy Degradation
en-us	96.08%	90.66%	(5.42)%
fr-fr	90.33%	85.95%	(4.38)%
es-es	92.84%	87.03%	(5.81)%
it-it	92.70%	89.78%	(2.92)%
de-de	92.35%	85.56%	(6.79)%
zh-cn	98.23%	94.43%	(3.80)%

Table 4. Test accuracy of 'universal' FIS models trained on all language data (second column), and 'universal' FIS models trained on all but the held-out language data (third column). The performance degradation with universal FIS models trained on all but the held-out data versus the universal FIS models trained on all language data is shown on the fourth column.

ter to unseen languages compared to randomly selecting some language specific FIS model. All of the FIS models, both universal and language specific models, were trained using the same parameter settings, i.e. same learning rate, number of trees, etc.

Experiment-3: Amount of Data Needed in a New Language to Train Universal Models: In this experiment, we tested the minimum amount of a new/unseen language training data required for training universal FIS model that would yield as good performance of as the universal FIS data that is trained as if substantial amount of unseen language is present at training time. For this experiment, we used the all unseen language training data but held-out one language data at first and then incrementally added back some percentage of held-out language to the training data. The goal is to find out how much training data would be sufficient if we would happen to adapt the universal FIS model to a specific language.

5.3. Results

Experiment-1: Table 3 presents results showing the accuracy of the selection models trained solely on that language's corpus and dialogue system by language specific FIS models. This is compared with a single, universal FIS model trained using all the languages. The en-us, es-es, zh-ch show the most gains among all languages, zh-ch having the most gain. The accuracy degraded for the remaining languages compared to the locale-specific FIS models. On average the language specific FIS models outperform the universal FIS models over six languages by about 0.6%. In examining the low performing universal FIS models compared to their language specific FIS models, it is noticeable that the distribution of the impact of features are different. For instance, in the language specific FIS model in one language the position matching features are far more important than explicit referring expression features, which was not as much important among universal FIS model features. It is possible that in some languages the user's might have preferred shorter utterances and use positional cues when referring to items on the screen. Further analysis is required to establish the likely cause. Nevertheless, the results indicate that the impact of features for language dependent FIS models are not identical across languages resulting in imbalance over some languages when trained globally.

Experiment-2: Table 4 presents results showing the accuracy of universal FIS models on previously unseen languages to which the SLU and FIS models are adapted. We observe that on average the absolute loss in accuracy for unseen languages compared to models trained when the language is observed is 4.82% in average. This indicates that the universal FIS model is able to achieve 85%-95% accuracy, which are fairly high making the FIS model usable with SDSs in different locales/languages.

To demonstrate the benefit of training a universal FIS model over simply reusing one of the existing language specific FIS models in a new/unseen language, Table 5 presents the FIS model accuracy when testing with each of the set of other language specific FIS model against the unseen language. The second column is just the accuracy of the universal FIS model where that language is previously unseen (copy over from Table 4 for ease of comparison). Every other column (column three to eight) is the accuracy of language specific FIS model on the unseen language data (row), where the ninth column is the average accuracy of all the language data (aveage of column three to column eight). The tenth column of Table 5 shows the average accuracy from all the other language specific FIS models on the unseen language (row). The last column is the absolute gain in accuracy by the universal FIS model trained with that language is held-out compared to the average of the rest of the language specific FIS models on the unseen language. As can be seen, the resulting gains from training a universal FIS model against language specific models lies between min. absolute value of +1.73% to max. absolute value of +11.59%, with an average

Held-Out Language	Universal FIS excludes held-out	Other language specific FIS models							Δ Gain in Accuracy by Universal FIS
		en-us	fr-fr	es-es	it-it	de-de	zh-cn	Average	
en-us	90.66%	–	87.87%	86.76%	88.09%	90.36%	87.87%	85.43%	+2.96%
fr-fr	85.95%	81.99%	–	82.39%	83.31%	82.73%	81.56%	82.40%	+3.55%
es-es	87.03%	86.98%	85.29%	–	86.73%	83.79%	83.74%	85.03%	+1.73%
it-it	89.78%	88.81%	85.11%	88.13%	–	85.96%	84.82%	86.67%	+3.21%
de-de	85.56%	84.52%	83.15%	83.02%	85.13%	–	81.04%	83.37%	+2.19%
zh-cn	94.43%	85.66%	85.09%	78.98%	80.03%	85.33%	–	82.84%	+11.59%

Table 5. Cross-testing of language specific FIS models on other (unseen/new) language FIS models. The last two columns are the average language dependent FIS accuracy of other languages on the unseen language and the gain in absolute accuracy of the universal FIS model against the individual models on unseen data.

of +4.2%.

Experiment-3: In experiment-2, we demonstrated that universal FIS models can be applied to unseen languages with minimal loss in accuracy. In this experiment we are searching for how much annotated data from the unseen language would be enough to get the same performance as if the unseen language is observed.

Our empirical analysis results are demonstrated in Figure 1. The x-axis is the % of the training data randomly selected from the unseen language and added to the training data that contains all other language data but that language. Once the percentage of unseen data is added to the overall training data, a new global FIS model is re-trained and tested on the held-out language's test data. We repeat this experiment by appending an additional 10% held-out language training data and re-train and report the numbers. Our goal is to find out at what percent of the unseen data will be enough to obtain as good performance as the overall universal FIS model trained on all other languages. The y-axis indicates the difference in accuracy of each of these new FIS models versus the overall universal FIS model that uses all the data from all languages we have in our corpora. Looking closely at Figure 1, for instance, when only 10% of en-us data is added to the overall FIS training data, the loss in accuracy compared to the overall universal FIS model is just ∼3%. It is to be noted that for some languages where there is more data (en-us, it-it) the elbow is around 20-30% (only 2% loss in accuracy) and for other languages where there is less training data the elbow is more towards 40% (around 3% loss in accuracy). As can be seen, with as little as 20% training data on a new language, the universal FIS models behave almost as good as the universal FIS models that includes the new language. Thus, universal FIS models can be scaled to new languages with even less effort in annotation in the new language.

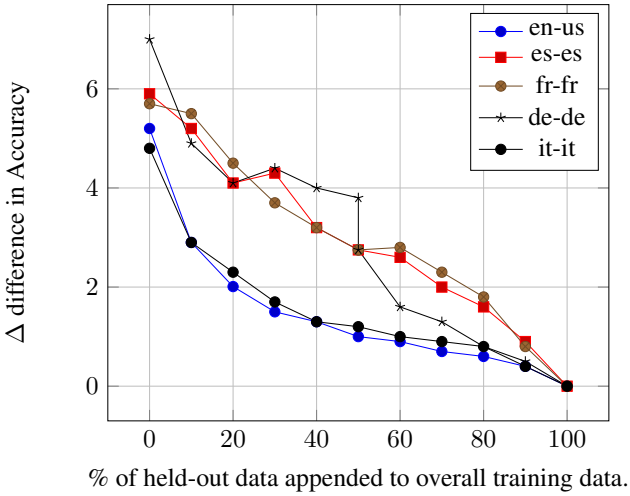


Fig. 1. Absolute Difference in Accuracy against universal FIS model trained on all data that includes all data from the held-our language. The delta is calculated based on the second column in Table 4.

6. CONCLUSIONS

This paper presented a universal FIS model for spoken dialogue systems deployed in multiple languages. We demonstrated that as the set of input features used by FIS models are largely language independent. A single, universal FIS model can be used in place of language specific FIS models with only a small loss in accuracy. In fact, the universal FIS model actually has an average gain of 0.6% (max of 3.59%) over language dependent FIS models. We also show that such an approach can generalize well to new unseen languages, with as low as 4.2% loss in accuracy when generalising to held-out (previously unseen) languages. The latter, which is achieved without retraining significantly, eases expansion of existing FIS models to new languages and avoids portability of existing models.

As a future work, we will investigate the performance of the universal FIS models on the overall dialog based on task completion metrics such as task success and user dissatisfaction.

7. REFERENCES

[1] A. Celikyilmaz, Z. Feizollahi, D. Hakkani-Tur, and R. Sarikaya, "Resolving referring expressions in conversational dialogs for natural user interfaces," *EMNLP*, 2014.

[2] R. De Mori, Frederic Bechet, D. Hakkani-Tur, M. McTear, G. Riccardi, and G. Tur, "Spoken language understanding: A survey," *Journal*, vol. 25, pp. 50–58, 2008.

[3] G. Tur, A. Celikyilmaz, and D. Hakkani-Tur, "Latent semantic modeling for slot filling in conversational understanding," *IEEE International Conference on Acoustics, Speech, and Signal Processing, Vancouver, Canada*, 2013.

[4] P. Xu and R. Sarikaya, "Targeted feature dropout for robust slot filling in natural language understanding," *ISCA - International Speech Communication Association*, 2014.

[5] R.A. Bolt, "Put-that-there: Voice and gesture at the graphics interface," *Computer Graphics*, 1980.

[6] L. Heck, D. Hakkani-Tur, M. Chinthakunta, G. Tur, R. Iyer, P. Parthasarathy, L. Stifelman, E. Shriberg, and A. Fidler, "Multi-modal conversational search and browse," *IEEE Workshop on Speech, Language and Audio in Multimedia*, 2013.

[7] M. Johnston, S. Bangalore, G. Vasireddy, A. Stent, P. Ehlen, M. Walker, S. Whittaker, and P. Maloor, "Match: an architecture for multimodal dialog systems," *ACL*, 2002.

[8] D. Huggins-Daines and A. I. Rudnicky, "Interactive asr error correction for touchscreen devices," *ACL Demo Session*, 2008.

[9] R Balchandran, M. E. Epstein, G. Potamianos, and L. Seredi, "A multi-modal spoken dialog system for interactive tv," *10th International Conference on Multimodal Interfaces*, 2008.

[10] D. Anastasiou, C.Jian, and D. Zhaekova, "Speech and gesture interaction in an ambient assited living lab," *1st Workshop on Speech and Multimodal Interaction in Assitive Environments at ACL'2012*, 2012.

[11] N. Pfleger and J. Alexandersson, "Towards resolving referring expressions by implicitly activated referents in practical dialog systems," *10th Workshop on the Semantics and Pragmatics of Dialog (SemDial-10)*, 2006.

[12] T. Tokunaga K. Funakoshi, M. Nakano and R. Iida, "A unified probabilistic approach to referring expressions," *Special Interest Group on Discourse and Dialog (SIGDIAL)*, 2012.

[13] D. Hakkani-Tur, A. Celikyilmaz, M. Slaney, and L. Heck, "Eye gaze for spoken language understanding in multi-modal conversational interactions," *International Conference on Multimodal Interaction*, 2014.

[14] M. Slaney A. Prokofieva and D. Hakkani-Tur, "Probabilistic features for connecting eye gaze to spoken language understanding," *ICASSP, IEEE - Institute of Electrical and Electronics Engineers*, 2015.

[15] T. Misu, R. Gupta A. Raux, and I. Lane, "Situated language understanding at 25 miles per hour," *SIGDIAL - Annual Meeting on Discourse and Dialogue*, 2014.

[16] P. Pantel, M. Gamon, and A. Fuxman, "Smart selection," *ACL*, 2014.

[17] V. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," *Doklady Akademii Nauk SSSR*, vol. 163, pp. 845–848, 1965.

[18] A. McCallum J. Lafferty and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," *ICML*, 2001.

[19] J. H. Friedman., "Greedy function approximation: A gradient boosting machiner," *Annals of Statistics*, 2001.

[20] V. Vapnik, ," *The nature of statistical learning theory*, 1995.

[21] C. Bishop, ," *Neural networks for Pattern recognition*, 1995.