Multi-Language Hypotheses Ranking And Domain Tracking for Open Domain Dialogue Systems

Paul A. Crook, Jean-Philippe Robichaud, Ruhi Sarikaya

Microsoft Corporation, Redmond WA 98052, USA

{pacrook, jerobich, ruhi.sarikaya}@microsoft.com

Abstract

Hypothesis ranking (HR) is an approach for improving the accuracy of both domain detection and tracking in multi-domain, multi-turn dialogue systems. This paper presents the results of applying a universal HR model to multiple dialogue systems, each of which are using a different language. It demonstrates that as the set of input features used by HR models are largely language independent a single, universal HR model can be used in place of language specific HR models with only a small loss in accuracy (average absolute gain of $+3.55\%$ versus $+4.54\%$), and also such a model can generalise well to new unseen languages, especially related languages (achieving an average absolute gain of $+2.8\%$ in domain accuracy on held out locales fr-fr, es-es, it-it; an average of 66% of the gain that could be achieve by training language specific HR models). That the latter is achieved without retraining significantly eases expansion of existing dialogue systems to new locales/languages.

Index Terms: dialogue systems, natural language understanding, hypothesis ranking, contextual domain classification, multi-language, locale expansion, language independence

1. Introduction

As natural language interaction, both spoken and typed, becomes mainstream across a range of devices, scaling the same applications and experiences to different locales and languages remains as a critical challenge.

Hypothesis Ranking (HR) was introduced previously [1] as a mechanism that improves the accuracy of a common architecture found in commercial multi-domain dialogue systems. Such systems typically first classify the user's utterance into one of the supported domains (or as an unsupported domain), this is followed by domain dependent intent classification and entity (slot) extraction. In such a set up the accuracy of domain classification is paramount as any errors made are significantly more noticeable as they tend to result in wildly incorrect system actions or responses. HR is a domain re-ranking mechanism within the dialogue manager stage of a dialogue system, *i.e.* post spoken language understanding (SLU), that benefits from having the full SLU domain, intent and slot analysis for all domains, as well as full session context and relevant back-end knowledge available to improve domain classification accuracy.

The input features to HR models are mostly derived features in the semantic space, *e.g.* the existence of a slot tag but not the actual words tagged, and are thus not language dependent. Thus, provided the set of domains handled by the dialogue systems are largely the same, HR models should generalise well across dialogue systems operating in different languages, including previously unseen languages. This, if true, is a useful property as it eases the expansion of such dialogue systems to new languages/locales.

1.1. Related Literature

Robichaud et al. [1] introduced the concept of Hypothesis Ranking (HR) for multi-domain dialogue systems and showed that ranking could produced significant gains in domain accuracy even when features were only extracted from the SLU's analysis. This work was only applied to an single language/locale (American English).

Various authors have applied ranking to SLU output, considering either n-best generated by using alternative ASR (automatic speech recognition) input to the SLU, alternative n-best generated by the SLU models or alternative SLU engines, *e.g.* Morbini et al., [2], Basili et al., [3], Ng and Lua [4], Dinarelli et al., [5, 6], Williams [7]. The closest cross-lingual approach, Dinarelli et al., [5, 6], used alternative languages for testing the effectiveness of their SLU re-ranking but did not attempted to train a cross-language model. This may have been due to each language corpus that was used, French, Italian and English, having been collected from a different domain. Their work also did not considered using a wider range of signals outside of those generated by the SLU, such as knowledge results or session context signals.

2. Hypothesis Ranking

The experimental dialogue system architecture is similar to that described in Robichaud et al. [1] and shown in Figure 1.

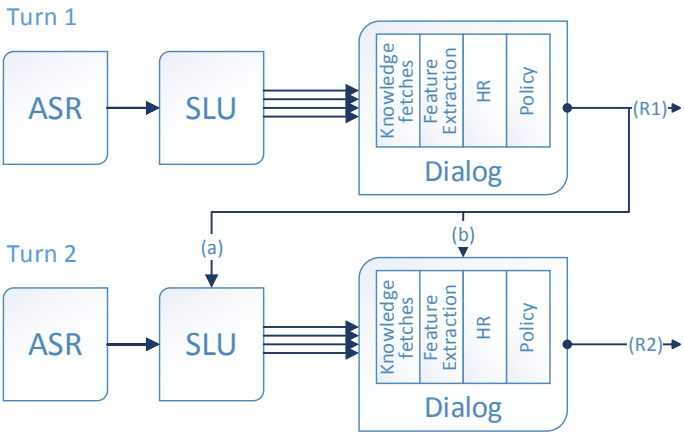

Figure 1: *Schematic diagram of the experimental spoken dialog system where (a) is a domain only contextual signal, (b) is the domain, intent and entities contextual signals, (R1) is selected result of turn 1, (R2) is selected result of turn 2.*

The language specific SLU module is a multi-turn, multi-domain statistical model that consists of a set of domain, intent and slot models. For each domain a *domain score* is generated

using support vector machine (SVM) models [8]. These domain models use the system's previous turn's selected domain as a contextual input signal which improves SLU domain prediction accuracy [9]. After domain classification, intents are then determined using a multi-class SVM intent model. Finally entities (slots) are tagged using conditional random fields (CRFs) sequence taggers [10]. One variation with the standard domain-intent-slot architecture is that in these experiments, for a given language, all of the SLU models for all of the domains are run in parallel as opposed to gating the running of the intent and slot models based on the domain prediction.

The output of the SLU is a set of semantic frames (SFs), one per domain, which contain intent and slot information, and associated scores. For each semantic frame relevant knowledge, *e.g.* database hits, are fetched and appended. These assemblies of SFs and knowledge results are referred to as *dialogue hypotheses*. Features are then extracted from these dialogue hypotheses which are used as input to the *hypotheses ranker (HR)*. The HR model, positioned as shown in Figure 1, is able to re-rank all of the domains recognised by the SLU based on a complete view of the SLU analysis, additionally back-end knowledge results and conversation context.

In this paper the HR models are Gradient Boosted Decision Tree (GBDT) models. Within a specific dialogue system a HR model assigns a score to each dialogue hypothesis, this score is then used to order the hypotheses. For training, each dialogue hypothesis is assigned a rating of 1 if its domain matches that selected by an annotator otherwise 0. The HR model score is then optimised using LambdaRank [11] to maximise the likelihood of it ranking a hypotheses with a rating of 1 in the top position.

Over 1,000 features are extracted for each hypothesis. These include binary features that indicate the presence or absence of a particular entity tag in that domain's analysis of the user's utterance, the domain's interpretation of the intent, the presence of canonicalised entities (not all tagged entities have a canonical form), coverage of tagged entities (as a percentage of the utterance length), *etc.* Other extracted features span the set hypotheses that are ranked together, *e.g.* whether a specific entity tag occurs anywhere in any of the hypotheses. Others are contextual features such as whether the hypothesis's domain matches the top ranked domain from the previous turn, how many entity tags a hypothesis has in common with the previous top ranked hypothesis, as well as the complete list of the previous turn's domains' scores. Features extracted from back-end domain knowledge include whether or not results can be generated for that hypothesis's combination of domain, intent and tagged entities.

None of the extracted features directly contain words or phrases from the user's utterance, *e.g.* no n-grams features. In setting up HR we deliberately decided to avoid using lexical features primarily to avoid the ranking model from recomputing the lower level lexical analysis already undertake by the SLU but also with an eye to portability between languages. Although some features are possibly influenced by the language in which the dialogue system is operating, for example the coverage of tagged entities as a percentage of the utterance length, all of the features can be computed in all languages. In addition none of the features directly indicate the language or locale of the dialogue system in which the HR model is deployed.

Our hypothesis is thus, *that given a consistent approach to SLU and back-end knowledge resources across all languages/locales, a single universal HR model shared between dialogue systems operating in different languages should be able to achieve accuracy close to that achievable by HR models trained specifically for each language. Furthermore, that such a universal HR model should generalise well to such a dialogue system operating in an unseen language/locale.*

2.1. Experiments

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. Roughly equal amounts of training data, around one hundred thousand utterances per language, were collected for six different languages-locales; French (fr-fr), German (de-de), Italian (it-it), Spanish (es-es), American English (en-us), and Chinese Mandarin (zh-cn). The corpora for all of the languages/locales span the same 9 distinct domains with multiple intents per domain.

Six dialogue systems were set up, one corresponding to each of the language-locale pairs for which user data was collected. Each of these dialogue systems has a different, language specific, SLU module and locale 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 corpora and systems were matched based on the language used – assuming either perfect language identification or that user will set the preferred language of the device and thus language identification is not required.

The training corpora were run through their corresponding language dialogue system just until the feature extraction stage in Figure 1. Features were then collected and stored from the set of hypothesis generated. For second and subsequent turns within a session, in the absence of an existing HR model, the contextual signals *(a)* and *(b)* were taken as being the domain, intent and entities contained in the previous turn hypothesis that had the highest SLU domain score. These signals are used as part of the captured hypothesis feature set. This is a-work-around to the bootstrapping issue that a HR model will, when in operation, effect the previous turn domain selection that it sees on subsequent turns. The result is a set of training examples with input features required by the HR models which are associated with human annotated domain labels as supervisory signals. A separate test corpora was also collected for each locale and processed in the same way. The test corpora was held out from SLU model training as well as HR model training.

The collected featurised data is used as an off-line training and test set for HR model training and testing in the following experiments. Accuracy of HR models is measured by comparing the top ranked hypothesis's domain with the annotated domain, and counting the percentage of matches. Similarly for the SLU, accuracy is measured by comparing the top scoring domain with the annotated domain.

Two experiments were run. In the first a single, universal HR model is trained on the complete 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 accuracy gain in domain prediction between the multi-turn aware SLU and the universal HR model for each system are recorded. These gains are then compared to those achieved by language specific HR models, each of which are trained solely on the language corpus that matches that dialogues system's language/locale.

In a second experiment one language's training data is completely held-out and a 'universal' HR model trained solely on the remaining languages. Its gain over the multi-turn aware SLU is then measured on the unseen language that has been held-out. This was repeated for all the languages. As a comparison each of the language specific HR models trained in the previous experiment were also tested against non-matching languages in order to test the assumption that a universal HR model would generalise better to unseen languages compared to randomly selecting some language specific HR model.

All of the HR models, both universal and language specific models, were trained using the same parameter settings, *i.e.* same learning rate, number of trees, *etc.* Thus, in principle, they have the same resolution power for learning the hypothesis scoring function. Parameter values used were values that had previously been found to generate good performance in language specific HR models. No parameter sweeping was undertaken to try and find optimal parameters for the universal HR model.

2.2. Results

Locale-language	Turns	Language specific HR gain	Uni. HR gain	Δ	% of specific HR gain
de-de	**All**	$\mathbf{+4.98\%}$	$\mathbf{+3.57\%}$	$\mathbf{-1.41\%}$	$\mathbf{71.7\%}$
	1^{st}	$+2.60\%$	$+1.79\%$	-0.81%	68.8%
	$2^{nd}+$	$+6.69\%$	$+4.85\%$	-1.84%	72.5%
fr-fr	**All**	$\mathbf{+6.29\%}$	$\mathbf{+5.70\%}$	$\mathbf{-0.59\%}$	$\mathbf{90.6\%}$
	1^{st}	$+9.04\%$	$+8.28\%$	-0.76%	91.6%
	$2^{nd}+$	$+4.22\%$	$+3.77\%$	-0.45%	89.3%
it-it	**All**	$\mathbf{+4.56\%}$	$\mathbf{+3.62\%}$	$\mathbf{-0.94\%}$	$\mathbf{79.4\%}$
	1^{st}	$+4.07\%$	$+3.01\%$	-1.06%	74.0%
	$2^{nd}+$	$+4.96\%$	$+4.08\%$	-0.88%	82.3%
es-es	**All**	$\mathbf{+2.95\%}$	$\mathbf{+2.64\%}$	$\mathbf{-0.31\%}$	$\mathbf{89.5\%}$
	1^{st}	$+1.68\%$	$+1.17\%$	-0.51%	69.6%
	$2^{nd}+$	$+3.69\%$	$+3.49\%$	-0.20%	94.6%
en-us	**All**	$\mathbf{+5.42\%}$	$\mathbf{+4.03\%}$	$\mathbf{-1.39\%}$	$\mathbf{74.4\%}$
	1^{st}	$+3.22\%$	$+1.98\%$	-1.24%	61.5%
	$2^{nd}+$	$+9.01\%$	$+7.35\%$	-1.66%	81.6%
zh-cn	**All**	$\mathbf{+3.05\%}$	$\mathbf{+1.71\%}$	$\mathbf{-1.34\%}$	$\mathbf{56.1\%}$
	1^{st}	$+3.79\%$	$+2.72\%$	-1.07%	71.8%
	$2^{nd}+$	$+2.68\%$	$+1.21\%$	-1.47%	45.1%

Table 1: Universal HR model trained on all languages versus HR Models train on one specific language. Gain is increase in domain accuracy compared to SLU domain selection. Last two columns report the delta and ratio between the universal and language specific HR models

Table 1 presents results showing the gain in domain accuracy achieved by language specific HR models trained solely on that language's corpus and dialogue system. This is compared with a single, universal HR model trained using all the languages and dialogue systems. Average gain is reported for all session turns (shown in bold) and then broken down into the average over only first turns of each session, and the average over only follow-up turns ($2^{nd}+$ turns). It is worth noting that the gain is measured with respect to a strong baseline in the form of a multi-turn SLU that is aware of the domain selected during the previous system turn, *e.g.* as in Xu and Sarikaya [9]). The HR gains would be larger for $2^{nd}+$ turns if tested against a non-contextual SLU; the average gain in domain accuracy on follow-up turns due to the context aware SLU is on average 6%.

Considering the average gain over all session turns, both the universal HR model and the language specific models demonstrate significant positive accuracy gains over the SLU domain prediction for all languages. The universal HR model has an average gain, computed over all languages and all turns, of $+3.55\%$. The language specific HR models have a combined average of $+4.54\%$. In all cases the language specific HR models outperform the universal HR model. However, the universal model achieves on average 77% of the locale specific model gain (averaged over all turns and locales). This corresponding to an average loss in accuracy of 1.0%.

Held-out Locale-Language	Turns	'Uni.' HR gain on unseen language	Ratio 'uni.' HR gain v. held-out lang. HR
de-de	**All**	$\mathbf{+1.66\%}$	$\mathbf{0.333}$
	1^{st}	$+0.07\%$	0.027
	$2^{nd}+$	$+2.81\%$	0.420
fr-fr	**All**	$\mathbf{+3.41\%}$	$\mathbf{0.542}$
	1^{st}	$+3.67\%$	0.406
	$2^{nd}+$	$+3.22\%$	0.763
it-it	**All**	$\mathbf{+2.70\%}$	$\mathbf{0.592}$
	1^{st}	$+1.70\%$	0.418
	$2^{nd}+$	$+3.48\%$	0.706
es-es	**All**	$\mathbf{+2.48\%}$	$\mathbf{0.841}$
	1^{st}	$+1.10\%$	0.655
	$2^{nd}+$	$+3.29\%$	0.892
en-us	**All**	$\mathbf{-2.95\%}$	$--$
	1^{st}	-3.24%	$--$
	$2^{nd}+$	-2.49%	$--$
zh-cn	**All**	$\mathbf{-0.08\%}$	$--$
	1^{st}	$+2.24\%$	0.591
	$2^{nd}+$	-1.25%	$--$

Table 2: Reporting gain of 'universal' HR models trained on all but one language and tested on the unseen language's test set. The 2nd column is the ratio of that 'universal' HR model's gain versus the gain achieved by a language specific HR model trained on the held-out language. No ratio is shown when the 'universal' model had negative gain

Table 2 presents results showing the principle benefit of training a universal HR model. That of being able to reliably apply the existing model to a previously unseen language to which a dialogue system is being adapted. The third column, *'Uni.' HR gain on unseen language*, presents the gain achieved by a universal HR model which has not been trained on that language. In nearly all cases the gain, averaged over all session turns, is positive and for closely related languages, *e.g.* latin languages fr-fr, it-it and es-se, the model generalises very well to the unseen member of this set. The one exception is for American English (en-us) where there is an overall loss of -2.95%. In examining the en-us corpus it is noticeable that it has a much lower number of user turns per session, around 1.61 turns/session, compared with the other languages where the number of user turns per session is between 2.47 and 3.25, with an average of 2.79 (excluding en-us). The distribution in domain usage is similar but not identical across languages. It is possible that en-us user's either prefer shorter tasks or are completing tasks in less turns – the earlier release in the en-us market make it possibly that users having become accustom to the dialogue interface. Further analysis is required to establish the likely cause.

A further comparison can be made between the performance of universal HR models that have not seen a particular

Held-out Locale-Language	Turns	Other language specific HR models							
		de-de	fr-fr	it-it	es-es	en-us	zh-cn	Avg. Gain	Max.
de-de	**All**	–	**+1.72%**	**+1.45%**	**+1.98%**	**+0.24%**	**−0.32%**	**+1.01%**	**+1.98%**
	1^{st}	–	+0.67%	+0.40%	+0.58%	−0.79%	−0.30%	+0.11%	+0.67%
	$2^{nd}+$	–	+2.48%	+2.21%	+2.98%	+0.99%	−0.33%	+1.67%	+2.98%
fr-fr	**All**	**+1.41%**	–	**+5.48%**	**+3.02%**	**+1.39%**	**+0.74%**	**+2.41%**	**+5.48%**
	1^{st}	+1.08%	–	+8.48%	+3.14%	+1.18%	+2.64%	+3.30%	+8.48%
	$2^{nd}+$	+1.65%	–	+3.23%	+2.93%	+1.54%	−0.69%	+1.73%	+3.23%
it-it	**All**	**+2.21%**	**+3.13%**	–	**+2.16%**	**+0.11%**	**+0.43%**	**+1.61%**	**+3.13%**
	1^{st}	+2.42%	+2.25%	–	+0.43%	−1.11%	+0.88%	+0.97%	+2.24%
	$2^{nd}+$	+2.04%	+3.81%	–	+3.49%	+1.04%	+0.08%	+2.09%	+3.81%
es-es	**All**	**+0.39%**	**+2.32%**	**+2.06%**	–	**−0.64%**	**−0.32%**	**+0.76%**	**+2.32%**
	1^{st}	+1.34%	+0.46%	+1.07%	–	−2.51%	−0.12%	+0.05%	+1.34%
	$2^{nd}+$	−0.17%	+3.41%	+2.64%	–	+0.45%	−0.44%	+1.18%	+3.41%
en-us	**All**	**−1.86%**	**−5.59%**	**−2.57%**	**+0.75%**	–	**−1.29%**	**−2.11%**	**+0.75%**
	1^{st}	−1.91%	−7.68%	−4.53%	+0.06%	–	−1.33%	−3.08%	+0.06%
	$2^{nd}+$	−1.80%	−2.18%	+0.63%	+1.87%	–	−1.22%	−0.54%	+1.87%
zh-cn	**All**	**+0.47%**	**−0.31%**	**−0.18%**	**−1.11%**	**+0.70%**	–	**+1.62%**	**+1.89%**
	1^{st}	+1.87%	+1.70%	+1.89%	+1.60%	+1.05%	–	+1.62%	+1.89%
	$2^{nd}+$	−0.24%	−1.32%	−1.22%	−2.47%	+0.53%	–	−0.94%	+0.53%

Table 3: Cross testing of language specific HR models on other language dialogue systems. Last two columns are the average and maximum gain of that row

language, and language specific models trained specifically on that language. Column four of Table 2 presents the ratio of the gain for each 'universal' HR model for which that language was never seen versus the HR model trained specifically on that language. In comparing these gains it can be seen that over the European languages-locales the universal HR that has not seen the language is achieving on average 0.57 of the gain of a model trained specifically on that language and that in one case, es-es, the ratio is 0.84, *i.e.* training a language specific es-es HR model only produces an additional half percent gain over a universal model that was not trained at all on es-es.

To demonstrate the benefit of training a universal model over simply reusing one of the existing language specific HR models for a dialogue system in a new language, Table 3 presents the HR model gains when testing with each of the set of other language specific model against the unseen language. The last two columns of Table 3 show the average and maximum gain for each row. As can be seen, the resulting gains from trying language specific HR models are some what unpredictable. Whereas, except for en-us and zh-cn, 'universal' HR models (from Table 2) typically generalises better and exceed the average of the gain of the other languages' HR models. For es-es, it-it and de-de their performance is close to or exceeds the maximum achieved by using other language models.

3. Discussion & Future work

These initial results are interesting in that when setting up a dialogue system in a new language the data required to train statistical models is often not available in sufficient quantities or quality. If some existing model can be deployed to that new language with some reasonable likelihood of performing well, given that there is little data available at that time to check the performance, this eases adoption of that technology. It also reduces the development and maintenance costs associated with locale expansion by reducing the number of models.

It is interesting to note the similarity in performance between the related language-locales suggest that the input features are still capturing cross language similarities, either in utterances or usage. We have not been at all selective in the set of input features used, preferring to rely on the ability of GBDT training to disregard features that are of little value. Nor have we explored the parameter space of GBDT-LambdaRank learning. The indiscriminate use of features may hurt generalisation of universal HR models to new languages, especially when, for example, those features only exist in that language. Thus another possible cause of the failure of the universal HR model, for which en-us training data was held-out, to generalize well to en-us could be related to the fact that the feature set used in all models was derived originally for en-us and thus possibly is overly specific to that language.

Similarly the parameter set used were *known-good* values for language specific HR models but possibly not the optimal values for training the universal HR models that are trained on 5-6 times the amount of data.

Thus we are working on trying to optimise the parameter values for both language specific and universal HR models and developing approaches for the exploration of the input feature set space in order to both promote generalisation and simultaneously close the gap between the universal and language specific HR models.

4. Conclusion

This paper presents the results of applying a universal HR model to dialogue systems that were built to operate in multiple languages. It demonstrates that as the set of input features used by HR models are largely language independent a single, universal HR model can be used in place of language specific HR models with only a small loss in accuracy. The universal HR model has an average gain of $+3.55\%$ over a multi-turn context aware SLU, while language specific HR models have a combined average of $+4.54\%$. We also show that such an approach can generalises well to new unseen languages, especially where those languages form similar language groups,*e.g.* average gain of $+2.8\%$ when generalising to held-out languages fr-fr, es-es or it-it. That the latter is achieved without retraining significantly eases expansion of existing dialogue systems to new locales/languages and avoids maintenance of multiple models.

5. References

[1] J.-P. Robichaud, P. A. Crook, P. Xu, O. Z. Khan, and R. Sarikaya, "Hypotheses ranking for robust domain classification and tracking in dialogue systems," in *Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH 2014)*, September 2014.

[2] F. Morbini, K. Audhkhasi, R. Artstein, M. V. Segbroeck, K. Sagae, P. S. Georgiou, D. R. Traum, and S. S. Narayanan, "A reranking approach for recognition and classification of speech input in conversational dialogue systems." in *IEEE Workshop on Spoken Language Technology (SLT)*, December 2012, pp. 49–54.

[3] R. Basili, E. Bastianelli, G. Castellucci, D. Nardi, and V. Perera, "Kernel-based discriminative re-ranking for spoken command understanding in hri," in *AI*IA 2013: Advances in Artificial Intelligence*, ser. Lecture Notes in Computer Science, M. Baldoni, C. Baroglio, G. Boella, and R. Micalizio, Eds. Springer International Publishing, 2013, vol. 8249, pp. 169–180.

[4] H.-I. Ng and K.-T. Lua, "Dialog input ranking in a multi-domain environment using transferable belief model," in *4th SIGDIAL Workshop on Discourse and Dialogue*, 2003.

[5] M. Dinarelli and S. Rosset, "Hypotheses selection criteria in a reranking framework for spoken language understanding," in *Conference of Empirical Methods for Natural Language Processing*, Jul 2011.

[6] M. Dinarelli, A. Moschitti, and G. Riccardi, "Discriminative reranking for spoken language understanding," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 2, pp. 526–539, February 2012.

[7] J. D. Williams, "Web-style ranking and slu combination for dialog state tracking," in *Proceedings of SIGDIAL*, June 2014.

[8] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[9] P. Xu and R. Sarikaya, "Contextual domain classification in spoken language understanding systems using recurrent neural network," in *Proceeding of ICASSP*, 2014.

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

[11] C. J. C. Burges, K. M. Svore, P. N. Bennett, A. Pastusiak, and Q. Wu, "Learning to Rank Using an Ensemble of Lambda-Gradient Models," *Journal of Machine Learning Research*, vol. 14, pp. 25–35, 2011.