UNSUPERVISED INDUCTION AND FILLING OF SEMANTIC SLOTS FOR SPOKEN DIALOGUE SYSTEMS USING FRAME-SEMANTIC PARSING

Yun-Nung Chen, William Yang Wang, and Alexander I. Rudnicky

School of Computer Science, Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213-3891, USA
{yvchen, yww, air}@cs.cmu.edu

ABSTRACT

Spoken dialogue systems typically use predefined semantic slots to parse users' natural language inputs into unified semantic representations. To define the slots, domain experts and professional annotators are often involved, and the cost can be expensive. In this paper, we ask the following question: given a collection of unlabeled raw audios, can we use the frame semantics theory to automatically induce and fill the semantic slots in an unsupervised fashion? To do this, we propose the use of a state-of-the-art frame-semantic parser, and a spectral clustering based slot ranking model that adapts the generic output of the parser to the target semantic space. Empirical experiments on a real-world spoken dialogue dataset show that the automatically induced semantic slots are in line with the reference slots created by domain experts: we observe a mean averaged precision of 69.36% using ASR-transcribed data. Our slot filling evaluations also indicate the promising future of this proposed approach.

Index Terms— Unsupervised slot induction, semantic slot filling, semantic representation.

1. INTRODUCTION

A number of recent and past efforts in industry (e.g. Google Now[1] and Apple's Siri[2]) and academia (e.g. [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]) have focused on developing semantic understanding techniques for building better spoken dialogue systems (SDS). The role of spoken language understanding (SLU) is of great significance to SDS: in order to capture the variation in language use from dialogue participants, the SLU component must create a mapping between the natural language inputs and a semantic representation that captures users' intentions.

As pointed out by Wang et al. [13], developing such a SLU-based interactive system can be very challenging: in the initial stage, domain experts or the developers themselves have to manually define the semantic frames, slots, and possible values that situates the domain-specific conversation scenarios. However, this approach might not generalize well to real-world users, and the predefined slot definition can be limited, and even bias the subsequent data collection and annotation. Another issue is about the efficiency: the manual definition and annotation process for domain-specific tasks can be very time-consuming, and have high financial costs. Finally, the maintenance cost is also non-trivial: when new conversational data comes in, developers, domain experts, and annotators have to manually analyze the audios or the transcriptions for updating and expanding the existing grammars.

Given a collection of unlabeled raw audio files, we investigate an unsupervised approach for automatic induction and filling of semantic slots. To do this, we use a state-of-the-art probabilistic frame-semantic parsing approach [14], and perform an unsupervised spectral clustering approach to adapt, rerank, and map the generic FrameNet style semantic parses to the target semantic space that is suitable for the domain-specific conversation settings [15]. To evaluate the performance of our approach, we compare the automatically induced semantic slots with the reference slots created by domain experts. Furthermore, we evaluate the accuracy of the slot filling (also known as form filling) task on a real-world SDS dataset, using the induced semantic slots. Empirical experiments show that the slot creation results generated by our approach aligns well with those of domain experts, and the slot filling experiment result suggests that our system is accurate in extracting the semantic information from users' input. Our main contributions of this paper are three-fold:

- We propose an unsupervised method for automatic induction and filling of semantic slots from unlabeled speech data, using probabilistic frame-semantic parsing;
- We show that the performance of this unsupervised slot induction method is in line with human-generated slots;
- We obtain interesting slot filling results that demonstrate the accuracy of our semantic information extraction system.

[1] http://www.google.com/landing/now/
[2] http://www.apple.com/ios/siri/

In the following sections, we outline related work in Section 2. We describe the proposed approach in Section 3. The experimental setup and results are shown in Section 4. Discussions are followed in Section 5, and we conclude in Section 6.

2. RELATED WORK

Early knowledge-based and statistical systems [3, 5, 8], for example, in the ATIS domain, all require developers to write syntactic and semantic grammars [16]. With the advent of statistical methods in language processing, Wong and Meng and Pargellis et al. are among the firsts to consider semiautomatic grammar induction [17, 18]. Their approaches are based on pure data-driven methods, which require considerable amount of manual efforts to clean up the output. There has been work in SLU that combines knowledge-based approach with the statistical approach. For example, Wang et al. have studied an HMM/CFG model that alleviates the need for massive human annotated grammar [16].

Unsupervised methods for automatic semantic slot induction that make use of the frame-semantic formalism from linguistic theories, on the other hand, have not been well studied in the SDS community. Although there has been work using semi-supervised learning approaches to explore unlabeled data [19], their focus was not on automatic induction and filling of slots. More broadly, Chung et al. and Wang et al. have studied the problem of obtaining in-domain natural language data for SDS semantic understanding, which address another important issue for building SDS with limited resources, at the other end of the spectrum [20, 13]. However, their domain-specific semantic representations are also predefined.

Despite our dialogue domain, our approach is relevant to ontology induction from text in the natural language understanding (NLU) community. For example, Poon and Domingos proposes an unsupervised ontology induction approach using Markov logic network [21]. Chambers and Jurafsky have studied unsupervised approaches for information extraction without templates [22]. Recently, Cheung et al. have investigated an HMM-like generative model to induce semantic frames in long passages across multiple documents [23]. Titov and Klementiev use a Bayesian approach to induce semantic roles [24], and Lang and Mirella have also proposed graph-based latent variable models for semantic role induction [25]. However, our SLU problem is different than the NLU problem: first, since spoken dialogue utterances are typically very short and noisy, they are often more syntactically and semantically ill-formed, as comparing to the newswire data that is commonly used in NLU problems. Also, since the SDS typically focus on specific domains, the resources for building such systems can be very limited. For example, the conventional semantic resources such as WordNet [26], were mostly built on top of written text data, so that it might not be *directly applicable* to the tasks in spoken language processing.

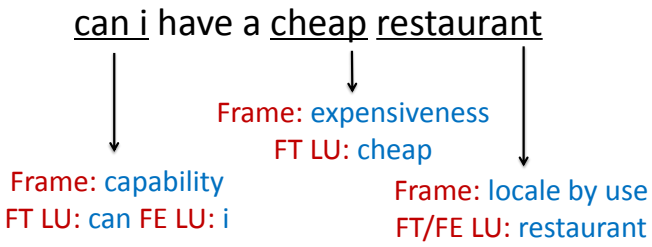

Fig. 1. An example of probabilistic frame-semantic parsing on ASR output. FT: frame target. FE: frame element. LU: lexical unit.

3. THE PROPOSED APPROACH

Our main motivation is to use a FrameNet-trained statistical probabilistic semantic parser [14] to generate initial frame-semantic parses from automatic speech recognition (ASR) decodings of the raw audio conversation files. Obviously, the question one might ask is: how do we map and adapt the FrameNet-style frame-semantic parses to the semantic slots in the target semantic space, so that they can be used practically in the spoken dialogue systems? To tackle this issue, we formulate the semantic mapping and adaptation problem as a reranking problem: we propose the use of an unsupervised spectral clustering based slot ranking model to rerank the list of most frequent parses from an unlabeled corpus.

In the remainder of the section, we introduce FrameNet and statistical semantic parsing in Section 3.1. Then we describe the slot ranking model that we use to adapt the generic semantic parsing outputs to target semantic space in Section 3.2.

3.1. Probabilistic Semantic Parsing

FrameNet is a linguistically-principled semantic resource that includes considerable annotations about predicate-argument semantics, and the associated lexical units in English [15]. FrameNet is developed based on a semantic theory called Frame Semantics [27]. The theory believes that the meaning of most words can be expressed on the basis of semantic frames, which is represented as three major components: *frame (F), frame elements (FE)*, and *lexical units (LU)*. For example, the frame "food" contains words referring to items of food. A descriptor frame element within the "food" frame indicates the characteristic of the food. For example, the phrase "`low fat milk`" should be analyzed with "`milk`" evoking the food frame and "`low fat`" filling the descriptor FE of that frame.

SEMAFOR is a state-of-the-art semantic parser for frame-semantic parsing [14, 28]. Trained on manually annotated

sentences in FrameNet, SEMAFOR is relatively accurate in predicting semantic frames, FE, and LU from raw text. Augmented by the dual decomposition techniques in decoding, SEMAFOR also produces the semantically-labeled output in a timely manner. Considering that SEMAFOR needs professional annotations and takes long time to train, the main contribution of the paper is to apply frame semantics theory to domain-specific tasks for dialogue systems in an unsupervised fashion.

In our approach, we parse all ASR-decoded utterances in our corpus using SEMAFOR, and we extract all frames from semantic parsing results as slot candidates, where the LUs that correspond to the frames are extracted for slot-filling. For example, in the Figure 1, we show an example of SEMAFOR parsing on an ASR-decoded text output. Thus here, SEMAFOR generates three frames (capability, expensiveness, and locale by use) for the utterance, which we consider as slot candidates. Note that for each slot candidate, SEMAFOR also includes the corresponding lexical unit (`can i`, `cheap`, and `restaurant`), which we consider as possible slot fillers.

Since SEMAFOR was trained on FrameNet annotation, which has a more generic frame-semantic context, not all the frames from the parsing results should be used as the actual slots in the domain-specific dialogue systems. For instance, in the example of Figure 1, we see that the "expensiveness" and "locale by use" frames are essentially the key slots for the purpose of dialog understanding of a SDS in the restaurant query domain, whereas the "capability" frame does not convey particular valuable information to SLU. In order to fix this issue, we compute the prominence of these slot candidates, use a slot ranking model to rerank the most frequent slots, and then generate a list of induced slots for the use in domain-specific dialogue systems.

3.2. Slot Ranking Model

The overall purpose of the ranking model is to distinguish generic semantic concepts and domain-specific concepts that were useful for spoken dialogue systems. To induce meaningful slots for the purpose of SDS, we compute the prominence of the slot candidates using a slot ranking model described below.

With the semantic parses from SEMAFOR, the model ranks the slot candidates by integrating two scores. One is the frequency of each candidate slot in the corpus, since slots with higher frequency may be more important. Another is the coherence of values the slot corresponds to. Because we assume that domain-specific concepts should focus on fewer topics and be similar to each other, the coherence of the corresponding values can help measure the prominence of the slots.

$$w(s_i) = \log f(s_i) + \alpha \cdot \log h(s_i), \tag{1}$$

where $w(s_i)$ is the ranking weight for the slot candidate s_i, $f(s_i)$ is the frequency of s_i from semantic parsing, $h(s_i)$ is the coherence measure of s_i, and α is the weighting parameter.

The coherence measure $h(s_i)$ is based on word-level clustering, where we cluster each word in the corpus using context information. For each slot s_i, we have the set of corresponding value vectors, $V(s_i) = \{\mathbf{v_1}, ..., \mathbf{v_j}, ..., \mathbf{v_J}\}$, where $\mathbf{v_j}$ is the j-th word vector constructed from the utterance including the slot s_i in the parsing results, and J is the number of these utterances. Then we have the set of the corresponding cluster vectors $C(s_i) = \{\mathbf{c_1}, ..., \mathbf{c_j}, ..., \mathbf{c_J}\}$. $\mathbf{c_j} = [c_{j1}, ..., c_{jk}, ..., c_{jK}]$, where c_{jk} is the frequency of words in $\mathbf{v_j}$ clustered into cluster k, and K is the number of clusters.

$$h(s_i) = \frac{\sum_{\mathbf{c_a}, \mathbf{c_b} \in \mathbf{C(s_i)}, \mathbf{c_a} \neq \mathbf{c_b}} \mathrm{Sim}(\mathbf{c_a}, \mathbf{c_b})}{|C(s_i)|^2}, \tag{2}$$

where $|C(s_i)|$ is the size of the set $C(s_i)$, $\mathrm{Sim}(\mathbf{c_a}, \mathbf{c_b})$ is the cosine similarity between cluster distribution vectors $\mathbf{c_a}, \mathbf{c_b}$, which are obtained from clustering approaches for each pair of values $\mathbf{v_a}$ and $\mathbf{v_b}$ s_i corresponds to. In general, a slot s_i with higher $h(s_i)$ usually focuses on fewer topics, which is more specific so that it is better for slots of dialogue systems.

For clustering, we formulate each word w as a feature vector $w = [r_1, r_2, ..., r_i, ...]$, where $r_i = 1$ when w occurs in the i-th utterance and $r_i = 0$ otherwise. With a set of word feature vectors, here we mainly use spectral clustering to cluster the words, because we assume that two words are topically-related when they occur in the same utterance. In spectral clustering [29], a key aspect is to define the Laplacian matrix L for generating the eigenvectors. Our spectral clustering approach can be summarized in the following five steps:

- **Calculate the distance matrix** $Dist$. Here, the goal is to compare the distance between each word pairs in the vector space. To do this, we use the the Euclidean distance as a metric, and use the K-nearest neighbor approach to select the top neighbors of each word.
- **Derive the affinity matrix** A. In order to convert the the distance matrix $Dist$ to a word affinity matrix, we apply heat kernel in order to account for irregularities in the data: $\theta_{ij} = \exp(-Dist^2/2\Sigma^2)$, where Σ is the variance parameter.
- **Generate the graph Laplacian** L. To generate the graph Laplacian L, we first define a diagonal degree matrix $D_{(i,i)} = \sum_{i=1}^{n} A_{(i,j)}$. Here, i and j are the row and column indices for the affinity matrix, and n is the dimension of the square matrix. So D is essentially the sum of all weighted connections of each word. Then, we define graph Laplacian as a symmetric normalized matrix $L = D^{-1/2} L D^{-1/2}$.
- **Eigendecomposition of** L. In the next step, we perform eigendecomposition of the graph Laplacian L,

and derive the eigenvectors V_{eigen} for the next clustering step.

- **Perform K-means clustering of eigenvectors V_{eigen}.** Finally, we normalize each row of V_{eigen} to be of unit length, and perform standard K-means clustering to obtain the cluster labels of each word.

The reason why we apply spectral clustering in this slot ranking model is because: 1) spectral clustering is very easy to implement, and can be solved efficiently by standard linear algebra techniques; 2) it is invariant to the shapes and densities of each cluster; 3) also, spectral clustering projects the manifolds within data into solvable space, and often outperform other clustering approaches. After spectral clustering, each word can have a cluster label c_k according to the clustering results.

4. EXPERIMENTS

To evaluate the effectiveness of our approach, we perform two evaluations. First, we examine the slot induction accuracy by comparing the reranked list of frame-semantic parsing induced slots with the reference slots created domain experts. Secondly, using the reranked list of induced slots and their associated slot fillers (value), we compare against the human annotation. For the slot-filling task, we evaluate both on ASR output of the raw audios, and the manual transcriptions.

4.1. Experimental Setup

In this experiment, we used the Cambridge University spoken language understanding corpus, which was also used several other SLU tasks in the past [30, 31]. The domain of the corpus is about restaurant recommendation in Cambridge, and the subjects of the corpus were asked to interact with multiple spoken dialogue systems for a number of dialogues in an in-car setting. There were multiple recording settings: 1) a stopped car with the air condition control on and off; 2) a driving condition; 3) and in a car simulator. The distribution of each condition in this corpus is uniform. An ASR system was used to transcribe the speech into text, and the word error rate was reported as 37%. The vocabulary size is 1868. The corpus contains a total number of 2,166 dialogues, resulting a total number of 15,453 utterances. The ratio of male vs. female is balanced, while there are slightly more native speakers than non-native speakers. The total number of slots in the corpus is 10, and they are: addr, area, food, name, phone, postcode, price range, signature, task, and type. We use develop set to tune the parameters α in (1) and the number of clusters K.

4.2. Slot Induction

To understand the accuracy of the induced slots, we measure the quality of induced slots as the proximity between induced slots and reference slots. We manually created a mapping table to indicate if the semantics of induced slots and reference slots are semantically related. This mapping table allows many-to-many mappings, and is shown in Table 1.

Table 1. The mapping table between induced slots (left) and reference slots (right)

Induced Slot	Reference Slot
speak on topic	addr
part orientational direction locale part inner outer	area
food origin	food
(NULL)	name
contacting	phone
sending	postcode
commerce scenario expensiveness range	price range
(NULL)	signature
seeking desiring locating	task
locale by use building	type

Table 2. The results of induced slots (%)

Approach	MAP	
	ASR	Manual
Frequency	67.31	59.41
K-Means	69.36	59.76
Spectral Clustering	**69.36**	**61.86**

For example, "commerce scenario $\rightarrow$ price range", "expensiveness $\rightarrow$ price", "food $\rightarrow$ food", and "direction $\rightarrow$ area" are the mappings between the induced slot and the reference slot. Since we define the adaptation task as a ranking problem, with a ranked list of induced slots, we can use the standard mean average precision (MAP) as our metric, where the induced slot is counted as correct when it has a mapping to a reference slot. Using the proposed spectral clustering approach, we observed the best MAP of 69.36 on the ASR output, and 61.86 on the manual transcription. The result is encouraging, because this indicates that a vast majority of the reference slots that are actually used in a real-world dialogue system can be induced automatically in an unsupervised fashion using our approach. One of the three methods we compared, spectral clustering performs best. The reason why ASR obtains better result than manual transcrip-

Table 3. The top-5 F1-measure slot-filling corresponding to matched slot mapping for ASR

SEMAFOR Slot	locale by use	speak on topic	expensiveness	origin	direction
Reference Slot	type	addr	price range	food	area
F1-Hard (%)	89.75	88.86	62.05	36.00	29.81
F1-Soft (%)	89.96	88.86	62.35	43.48	29.81

Table 4. The results of induced slots and corresponding values (%)

Approach	MAP-F1-Hard		MAP-F1-Soft	
	ASR	Manual	ASR	Mannual
Frequency	26.96	27.84	27.29	28.68
K-Means	27.38	27.99	27.67	28.83
Spectral Clustering	**30.52**	**28.40**	**30.85**	**29.22**

tion in this task may be that generic words might have higher WER, probably because users tend to speak keywords clearer than generic words. Therefore, the same generic word might correspond to different SEMAFOR parses in the ASR output due to recognition errors, and then these slot candidates will be ranked lower by the coherence model, which aligns to our initial goal: ranking domain-specific concepts higher and generic concepts lower. The detailed results are shown in the Table 2.

4.3. Slot Filling

While semantic slot induction is essential for providing semantic categories and imposing semantic constraints, we are also interested in understanding the performance of our induced slot fillers.

For each matched mapping between the induced slot and the reference slot, we can compute F-measure by comparing the lists of extracted slot fillers with the induced slots, and the slot fillers in the reference list. Considering that the slot fillers may contain multiple words, we have two ways to define whether two slot fillers match each other: hard matching and soft matching, where "hard" means that the values of two slot fillers should be exactly the same, and "soft" means that if the two slot fillers both contain at least one overlapping words, we count this comparison as a matched case. We show the top-5 hard-matching and soft-matching results of ASR in Table 3.

Since the MAP score can be weighted by either soft or hard matching, we compute the corresponding MAP-F1-Hard and MAP-F1-Soft scores, which evaluate the accuracy of both slot induction and slot-filling tasks together. The results of slot ranking model are shown in Table 4.

5. DISCUSSIONS

In the slot induction experiment, we observed some interesting induced slots: for example, "direction $\rightarrow$ area" and "expensiveness $\rightarrow$ price range". Here, "direction" and "expensiveness" are induced slots, whose semantic meanings correspond to "area" and "price range" in the reference slots respectively, which are essential to the task of SLU. Our best system obtains an MAP of 69.36, indicating that our proposed approach generates a good coverage of the domain-specific semantic slots for real-world SDS.

In the slot filling evaluation, although the overall F-measure is much lower than the slot induction task, it is pretty much expected: when the induced slot mismatch the reference slot, all the slot fillers will be judged as incorrect fillers. However, even though our dataset is different, the overall F-measure performance aligns with related work of template induction and slot filling in newswire based NLU tasks [22]. In addition, the best scoring systems in the past NIST slot filling evaluation also have a F-measure ≈ 0.3 [32, 33], indicating the challenging nature of the slot filling task. While we work in the SLU domain, it is entirely possible to apply our approach to the text-based NLU and slot filling tasks.

6. CONCLUSION

In this paper, we propose an unsupervised approach for automatic induction and filling of slots. Our work makes use of a state-of-the-art semantic parser, and adapts the linguistically-principled generic FrameNet-style outputs to the target semantic space that corresponds to a domain-specific SDS setting. In our experiments, we show that our automatically induced semantic slots align well with the reference slots, which are created by domain experts. In addition, we also study the slot-filling tasks that extract the slot-filler information from those automatically induced slots. In the future, we plan to increase the coverage of frame semantic parses by incorporating the WordNet semantic resource.

7. ACKNOWLEDGEMENT

The authors would like to thank Brendan O'Connor, Nathan Schneider, Sam Thomson, Dipanjan Das, and the anonymous reviewers for valuable comments.

8. REFERENCES

[1] Patti Price, "Evaluation of spoken language systems: The ATIS domain," in *Proceedings of the Third DARPA Speech and Natural Language Workshop*. Morgan Kaufmann, 1990, pp. 91–95.

[2] Roberto Pieraccini, Evelyne Tzoukermann, Zakhar Gorelov, J-L Gauvain, Esther Levin, C-H Lee, and Jay G Wilpon, "A speech understanding system based on statistical representation of semantics," in *Proc. of ICASSP*. IEEE, 1992, vol. 1, pp. 193–196.

[3] John Dowding, Jean Mark Gawron, Doug Appelt, John Bear, Lynn Cherny, Robert Moore, and Douglas Moran, "Gemini: A natural language system for spoken-language understanding," in *Proc. of ACL*. Association for Computational Linguistics, 1993, pp. 54–61.

[4] Scott Miller, Robert Bobrow, Robert Ingria, and Richard Schwartz, "Hidden understanding models of natural language," in *Proc. of ACL*. Association for Computational Linguistics, 1994, pp. 25–32.

[5] Wayne Ward and Sunil Issar, "Recent improvements in the CMU spoken language understanding system," in *Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics, 1994, pp. 213–216.

[6] Staffan Larsson and David R Traum, "Information state and dialogue management in the TRINDI dialogue move engine toolkit," *Natural language engineering*, vol. 6, no. 3&4, pp. 323–340, 2000.

[7] Wei Xu and Alexander I Rudnicky, "Task-based dialog management using an agenda," in *Proceedings of the 2000 ANLP/NAACL Workshop on Conversational systems-Volume 3*. Association for Computational Linguistics, 2000, pp. 42–47.

[8] Stephanie Seneff, "TINA: A natural language system for spoken language applications," *Computational linguistics*, vol. 18, no. 1, pp. 61–86, 1992.

[9] James F Allen, Donna K Byron, Myroslava Dzikovska, George Ferguson, Lucian Galescu, and Amanda Stent, "Toward conversational human-computer interaction," *AI magazine*, vol. 22, no. 4, pp. 27, 2001.

[10] Oliver Lemon, Kallirroi Georgila, James Henderson, and Matthew Stuttle, "An isu dialogue system exhibiting reinforcement learning of dialogue policies: generic slot-filling in the talk in-car system," in *Proc. of EACL*. Association for Computational Linguistics, 2006, pp. 119–122.

[11] Narendra Gupta, Gokhan Tur, Dilek Hakkani-Tur, Srinivas Bangalore, Giuseppe Riccardi, and Mazin Gilbert, "The at&t spoken language understanding system," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 1, pp. 213–222, 2006.

[12] Dan Bohus and Alexander I Rudnicky, "The RavenClaw dialog management framework: Architecture and systems," *Computer Speech & Language*, vol. 23, no. 3, pp. 332–361, 2009.

[13] William Yang Wang, Dan Bohus, Ece Kamar, and Eric Horvitz, "Crowdsourcing the acquisition of natural language corpora: Methods and observations," in *Proc. of the IEEE SLT 2012*, Miami, Florida, Dec. 2012, IEEE.

[14] Dipanjan Das, Nathan Schneider, Desai Chen, and Noah A Smith, "Probabilistic frame-semantic parsing," in *Proc. of HLT-NAACL*. Association for Computational Linguistics, 2010, pp. 948–956.

[15] Collin F Baker, Charles J Fillmore, and John B Lowe, "The Berkeley FrameNet project," in *Proc. of COLING*. Association for Computational Linguistics, 1998, pp. 86–90.

[16] Ye-Yi Wang, Li Deng, and Alex Acero, "Spoken language understanding," *Signal Processing Magazine, IEEE*, vol. 22, no. 5, pp. 16–31, 2005.

[17] Chin-Chung Wong and Helen Meng, "Improvements on a semi-automatic grammar induction framework," in *Proc. of ASRU*. IEEE, 2001, pp. 288–291.

[18] Andrew Pargellis, Eric Fosler-Lussier, Alexandros Potamianos, and Chin-Hui Lee, "A comparison of four metrics for auto-inducing semantic classes," in *Proc. of ASRU*. IEEE, 2001, pp. 218–221.

[19] Gokhan Tur, Dilek Hakkani-Tür, and Robert E Schapire, "Combining active and semi-supervised learning for spoken language understanding," *Speech Communication*, vol. 45, no. 2, pp. 171–186, 2005.

[20] Grace Chung, Stephanie Seneff, and Chao Wang, "Automatic induction of language model data for a spoken dialogue system," in *Proc. of SIGDIAL*, 2005.

[21] Hoifung Poon and Pedro Domingos, "Unsupervised ontology induction from text," in *Proc. of ACL*. Association for Computational Linguistics, 2010, pp. 296–305.

[22] Nathanael Chambers and Dan Jurafsky, "Template-based information extraction without the templates.," in *ACL*, 2011, pp. 976–986.

[23] Jackie Chi Kit Cheung, Hoifung Poon, and Lucy Vanderwende, "Probabilistic frame induction," *Proc. of HLT-NAACL*, 2013.

[24] Ivan Titov and Alexandre Klementiev, "A Bayesian approach to unsupervised semantic role induction," in *Proc. of ACL*. Association for Computational Linguistics, 2012, pp. 12–22.

[25] Joel Lang and Mirella Lapata, "Unsupervised induction of semantic roles," in *Proc. of HLT-NAACL*. Association for Computational Linguistics, 2010, pp. 939–947.

[26] George A Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[27] Charles J Fillmore, "Frame semantics and the nature of language," *Annals of the NYAS*, vol. 280, no. 1, pp. 20–32, 1976.

[28] Dipanjan Das, Desai Chen, André F. T. Martins, Nathan Schneider, and Noah A. Smith, "Frame-semantic parsing," *Computational Linguistics*, 2013.

[29] Ulrike Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[30] Matthew Henderson, Milica Gasic, Blaise Thomson, Pirros Tsiakoulis, Kai Yu, and Steve Young, "Discriminative spoken language understanding using word confusion networks," in *Proc. of SLT*. IEEE, 2012, pp. 176–181.

[31] Yun-Nung Chen, William Yang Wang, and Alexander I. Rudnicky, "An empirical investigation of sparse log-linear models for improved dialogue act classification," in *Proc. of ICASSP*, May 2013.

[32] Ang Sun, Ralph Grishman, Wei Xu, and Bonan Min, "New york university 2011 system for KBP slot filling," in *Proc. of TAC*, 2011.

[33] Mihai Surdeanu, Sonal Gupta, John Bauer, David Mc-Closky, Angel X Chang, Valentin I Spitkovsky, and Christopher D Manning, "Stanford's distantly-supervised slot-filling system," in *Proc. of TAC*, 2011.