



# Zero-shot semantic parser for spoken language understanding

Emmanuel Ferreira, Bassam Jabaian and Fabrice Lefèvre\*

CERI-LIA, University of Avignon, Avignon - France

firstname.lastname@univ-avignon.fr

## Abstract

Machine learning algorithms are now common in the state-of-the-art spoken language understanding models. But to reach good performance they must be trained on a potentially large amount of data which are not available for a variety of tasks and languages of interest. In this work, we present a novel zero-shot learning method, based on word embeddings, allowing to derive a full semantic parser for spoken language understanding.

No annotated in-context data are needed, the ontological description of the target domain and generic word embedding features (learned from freely available general domain data) suffice to derive the model. Two versions are studied with respect to how the model parameters and decoding step are handled, including an extension of the proposed approach in the context of conditional random fields. We show that this model, with very little supervision, can reach instantly performance comparable to those obtained by either state-of-the-art carefully handcrafted rule-based or trained statistical models for extraction of dialog acts on the Dialog State Tracking test datasets (DSTC2 and 3).

**Index Terms:** spoken language understanding, word embedding, zero-shot learning, out-of-domain training data.

## 1. Introduction

In dialogue systems, the Spoken Language Understanding (SLU) module has an intermediate role between the Automatic Speech Recognizer and the Dialogue Manager. Its role is to extract a list of semantic concept hypotheses from an input sentence transcription of the user’s query. Currently, the state-of-the-art SLU systems are based on probabilistic approaches and trained with various machine learning methods to tag the user input with these semantic concepts.

Dealing with supervised machine learning techniques requires a large number of sentences which are semantically annotated by humans. Annotated corpora are expensive (in cost and elaboration time) and domain dependent. Several studies compared probabilistic methods to train a SLU model [1, 2, 3]. Despite their good performance, these approaches are highly data dependent and thus hard to generalise.

To deal with this limitation, some research proposed an unsupervised annotation process (for instance based on latent Dirichlet allocation, e.g. [4]), others focused on the use of unsupervised [5, 6] or lightly supervised [7, 8] training approaches to cope with the lack of annotated resources by exploiting the semantic web for mining additional training data and enriching classification features.

Another group of studies tries to minimize the time of transcribing and collecting training data. While in [9] and [10] the

authors proposed to construct a tiny corpus to build a pilot system that is used in successive data collections, others employed active learning to reduce the time required for corpus annotation and verification [11, 12]. Other works proposed to reduce the cost of data collection by porting a system across language [13, 14] and domain [15, 16].

Furthermore, Dauphin et al. [17] proposed a zero-shot learning algorithm for Semantic Utterance Classification (SUC). This method tries to find a sentence-wise link between categories and utterances in a semantic space. A deep neural network can be trained on a large amount of unannotated and unstructured data to learn this semantic space.

In the same line of [18], in this work, we present a zero-shot learning method for SLU based on word embeddings. This approach requires neither annotated data nor in-context data. Indeed, only the ontological description of the target domain and generic word embedding features (learned from freely available and general purpose data) are required to obtain the model. Two versions are studied with respect to how the model parameters and decoding step are handled, including an extension of the proposed approach in the context of Conditional Random Fields. It is shown on the Dialog State Tracking (DSTC2 and DSTC3) testbeds [19, 20] that the proposed technique offers immediate performance comparable to those obtained with either carefully handcrafted rules-based or state-of-the-art trained models.

In Section 2 we describe the SLU task. Section 3 presents the proposed zero-shot learning approach followed by the presentation of some related works in Section 4. We present our experiments in Section 5 before some conclusions.

## 2. Spoken Language Understanding

The aim of the SLU module is to extract from a user utterance of  $n$  words,  $W = w_1, w_2, \dots, w_n$ , a valid sequence of  $m$  concept  $C = c_1, c_2, \dots, c_m$ , where  $c_i$  is a slot-value pair such as *food=Italian* or *destination=Boston*. In this paper, the semantics are based on the standards defined during the challenges embodied in the DSTC2 and DSTC3 corpora [19] wherein the extracted sequence of concepts is expressed as a sequence of Dialogue Acts (DAs) of the form *actype(slot=value)*.

The *actype* are task-independent and can be divided into 4 groups: information providing (*inform*), query (*request, requests, reqmore*), confirmation (*confirm, affirm, negate, deny*) and housekeeping (*hello, thankyou, bye, ...*). *Slots* and *values* are domain dependent and correspond to specific entries in the backend database. For instance the utterance “hello i am looking for a french restaurant in the south part of town” corresponds to the dialogue act sequence “*hello()*, *inform(food=french)*, *inform(area=south)*”.

Thus, the considered SLU task is a sequential tagging problem where possible tags are all task-specific *actype(slot=value)*

This work is partially supported by the ANR funded MaRDi project (ANR-12-CORD-0021). <http://mardi.metz.supelec.fr>.

combinations based on a pre-defined inventory of acttypes, slots and associated values and where each tag is associated to a particular chunk of words in  $W$ .

Due to their good performance, authors of [21] proposed to use Conditional Random Fields (CRF) ([22]) for language understanding and they are now largely used as a state-of-the-art model for this task. More recent approaches such as Deep Belief Network (DBN) [3], Recurrent Neural Network (RNN) [23, 24] or Recurrent Conditional Random Fields (R-CRF) [25, 26] yield state-of-the-art results on different SLU tasks. However, experimental studies show that their performance are close to those obtained with CRF.

### 3. Zero-shot learning for Spoken Language Understanding

The Zero-shot learning, as proposed in [27], corresponds to a learning situation where possible values for a class  $Y$  include cases that have been omitted from the training examples. In this study, we especially examine the problem of predicting the true semantic tag sequence of a user query without having seen any example of in-domain user utterances and thus in-context semantic tags.

In the same line, our proposition makes the assumption that no in-context data are available except the domain ontology description, denoted as Knowledge Base ( $K$ ). To exploit this limited and non-contextual resource, we use a feature space  $F$ , based on word embeddings, in a SLU parser to extract scored graphs of semantic tag sequence hypotheses from user utterances.

#### 3.1. Knowledge Base

The knowledge base  $K$  is a set of  $M$  examples:  $\{ex_i, y_i\}_{i=1:M}$  where  $ex_i$  is a lexicalisation of all the task-specific semantic of the ontological description and  $y_i$  is the associated class label from the set  $Y$ .

We limited ourselves to situations where an initial transformation can be obtained rather automatically by exploiting the task-related database and some generic dialogue management knowledge (e.g. possible user *acttypes*). However some techniques, as those presented in [11, 12], can also be exploited to gather some additional mappings in an unsupervised manner. So, to each semantic tag *acttype(slot=value)* is associated some related Out-Of-Context Surface Forms (OOCsFs).

For user dialogue acts (*acttype*), a set of lexical examples can be easily constituted manually. For instance “yes, ok, all right” for the *affirm* act or “what’s the SLOT” for *request*. Concerning slot and value pairs, their respective field’s name and value in the task related database are exploited (e.g. “food” for the slot food, “french” for the food value french). In order to extract the valid OOCsFs for the whole semantic tags (i.e. *acttype(slot=value)*) the previously described lexical forms are combined (e.g. “what’s the food?” for *request(food)*, or “is it moderate price range?” for *confirm(pricerange=moderate)*).

Even if  $K$  can be used to train a stochastic model such as CRF<sup>1</sup>, the resulting model suffers from two main limitations: on the one hand the vocabulary coverage of the training data is limited, so the number of Out-Of-Vocabulary (OOV) words in test can be very high and on the other hand the context is poorly handled since the sequentiality of the tags is not handled

<sup>1</sup>Since each tuple  $\{ex_i, y_i\}$  can be seen as a training sentence composed of  $w_1, \dots, w_n$  word tagged by their corresponding  $y_i$  semantic class

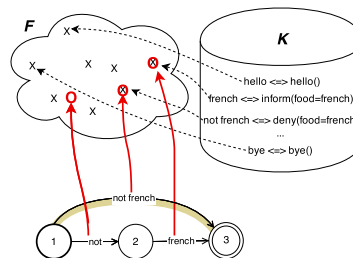


Figure 1: Illustration of the zero-shot learning SLU parsing.

in OOCsFs. To deal with these limitations a word-embedding-based semantic feature space  $F$  based on word embeddings has been employed.

#### 3.2. Word-embedding-based Semantic Feature Space

Recent progresses have been made in word embedding neural network learning [28, 29]. Several researches pinpointed the interest of exploiting some regularities between syntactic/semantic features of words and their corresponding embedding for different NLP tasks [30, 31]. Considering this representation avoids any task dependency as it is learnt on a wide coverage dataset (several billions of words). Moreover, some processes allow to adapt/transfer the model to specific task/language [32].

Since the goal of SLU is to extract some task-specific semantic information from imprecise word inputs conveyed by the user, we consider that a continuous representation of word is a natural generalisation mechanism which the system can leverage with. Even more, for a model trained on a knowledge base with a limited vocabulary, this mechanism can be an efficient solution to deal with both the important number of the OOV words and the sentence segmentation problem (in-context data) during decoding.

For that, we consider two methods to map words in the user utterance in the feature space  $F$ : In the first method, each OOV word is replaced by its  $n$  nearest known words (from  $K$ ) according to the semantic metric space  $F$  (e.g. cosine similarity). In the second method, illustrated in Figure 1, all possible contiguous word sequences (chunks) present in the user utterance are mapped to the known  $ex_i$ s present in  $K$ . For example if the user says “not french”, 3 different chunks are considered: “not”, “french” and “not french”. Then, these chunks are mapped to the feature space  $F$  with the same merging strategy used to map OOCsFs. The resulting points (red circles in Fig. 1) are then compared in terms of similarity to the known knowledge base entries (black crosses in Fig. 1).

In the two considered methods a sentence is not seen as a sequence of words anymore but as a graph.

#### 3.3. SLU parsing: extension to Conditional Random Fields

A graphical decoding has been proposed by [33] in order to obtain an efficient CRF-based translation system. This model is based on finite state transducers in which the different stages process are represented and can be composed all together. This proposition has been revisited by [34] for SLU allowing to consider the semantic tagging of a sentence as a composition of transducers in the following order:

$$\lambda_{understanding} = \lambda_S \circ \lambda_T \circ \lambda_F \quad (1)$$

Where  $\lambda_S$  is the acceptor of the source sentence  $s$ ;  $\lambda_T$  is a dictionary of tuples, combining sequences of the words and

their possible tags based on the tuples inventory; and  $\lambda_F$  is a feature matcher, which assigns probability scores to tuples using a trained model.

A similar architecture is adopted here.  $K$  is employed to derive our  $\lambda_T$ . Indeed, valid tuples are extracted by the mean of a k-nearest neighbors method which maps each chunk in the input sentence to some tuples  $(ex_i, y_i)$  from  $K$  according to some metric on  $F$  (e.g. cosine distance to the class-points). These tuples are considered as the edges of the  $\lambda_T$  transducer. Two configurations of the parser are employed depending on the nature of the chunks extracted from  $K$ . Indeed, the chunks (input/output) can be single words or sequence of  $n$  words (resp. denoted as word-parser and chunk-parser in the rest of the paper).

A statistical model (CRF in our study) can be trained even on words or on chunks in order to be used as  $\lambda_F$ . The best semantic sequence hypothesis at the utterance level is obtained by a best-path decoding on the finite-state machine (highlighted path in Fig. 1).

## 4. Related work

The problem of zero-shot learning has been addressed in the machine learning community for the last 5 years. The Larochelle’s early work [35] introduced the problem of zero-data learning for a character recognition problem. This work discerned two ways to tackle this issue. In the first, the input of the learning problem is connected with the corresponding class description and then a standard supervised learning algorithm can be used to train a model. While in the second, the model for a class is directly a function of its description.

In parallel, the authors of [27] defined the notion of semantic output code by proposing a zero-shot learning algorithm to predicts classes omitted from the training set. This algorithm also used a knowledge base of semantic properties of the known classes to explore novel classes. It was applied in the context of decoding novel thoughts from the neural activity of a person.

As an application of zero-shot learning to NLP our approach borrows from the proposition of [17] with several major differences: First with regard to how the semantic space is modeled, since no domain data is required in our proposition. Second in the task itself as we address complete semantic annotation of a sentence (SLU) and not whole utterance classification (SUC). With the same objective of minimising the need for costly training data, different approaches have been already applied for exploiting the semantic web for the SUC task. For example, authors of [36] proposed an unsupervised training approach for understanding systems based on the use of semantic knowledge from the Semantic Web. These propositions are based on a combination of web search retrieval and syntax-based dependency parsing. Anastasakos and Deoras [37] also exploited word embedding, though they proposed an approach to obtain task and domain specific embeddings to train an understanding system using an unsupervised training algorithm. They also proposed to transfer these embeddings from one language to another enabling training of a multilingual spoken language understanding system.

## 5. Experiments and Results

### 5.1. Data description

Experiments presented in this paper are based on the DSTC2 and DSTC3 datasets [19, 20]. Even if these research challenges

focused on tracking the user’s goal all along the dialogue (not only SLU), here we only exploit the fully annotated data (e.g. transcriptions, dialogue-act semantics) as a test set to evaluate our zero-shot semantic decoding approach on two realistic dialogue setups.

The DSTC2 challenge covers the domain of restaurant search. The SLU dataset of this challenge is composed of 11677 user utterances for train and 9890 user utterances for test. The DSTC3 extends the previous domain to also cover tourist information. The SLU data of this corpus is composed of 85 seed user utterances (10 dialogues) and 18715 for test.

### 5.2. Evaluation of the proposed model

The task-dependent knowledge bases employed in our experiments are derived from the two challenge ontologies, as well as from a shared generic dialogue information following the automatic procedure described in Section 3.1. For DSTC2 the semantic of the domain is represented by 8 slots and 215 values and for DSTC3, by 13 slots and 279 values. For both tasks 16 different act types are considered, resulting in 663 different semantic tags for DSTC2 and 855 for DSTC3. The OOCFSs (53) used to model act types were manually written and are shared across tasks (e.g. “say again” for the repeat act). In the two considered ontological descriptions, slots and values have meaningful (lexicalised) names and so they can be directly used as an OOCFS (e.g. “address”, “french”, “has tv”). Overall, 4160 automatically generated chunks are considered for DSTC2 (resp. 6555 for DSTC3).

In order to constitute the semantic space, a 300-dimensional word2vec [28] word-embedding model is considered. It has been trained on a large amount of wide coverage and freely available english corpora<sup>2</sup> with the Skip-gram algorithm (with a 10-words window). The resulting model is expected to exhibit some linguistic regularities as those showed in [38] as well as a linear structure that makes it possible to meaningfully combine the words by an element-wise addition of their word embeddings [39]. The latter technique is employed to directly map word chunks from both user inputs and OOCFSs to their corresponding word2vec representation by applying the sum of individual word representations.

Due to fact that word2vec behaves nicely with cosine similarity in the literature [28, 38] this metric is used in the nearest neighbors strategy to find related tuples for  $\lambda_T$  (see section 3.3). In order to obtain the feature matcher  $\lambda_F$  a CRF model is trained with wapiti [40] even on words or on chunks extracted from  $K$  (resp. for word-parser and chunk-parser).

We consider a CRF model<sup>3</sup> trained on the whole train dataset for each task as baseline to compare the performance of our proposition (denoted as CRF-train in Table 1). Notice that due to the fact that  $acttype(slot = value)$  semantic tags are not aligned with words in the considered corpora and since a word level tagging is a prerequisite to use the CRF model, we choose to use an adapted unsupervised alignment procedure following [41].

The results presented in Table 1 show that, with no training data, the chunk-parser gives better performance than the word-parser in terms of F-score (0.786 vs. 0.679 for DSTC2 and 0.817 vs. 0.552 for DSTC3). We show also that even that this approach has a lower performance than a CRF model trained on

<sup>2</sup>enwik9, One Billion Word Language Modelling Benchmark, the Brown corpus, English GigaWord from 1 to 5

<sup>3</sup>using bigram and unigram features with a neighborhood window of length 2 around the current word

Task	Model	F-score	P	R
DSTC2	CRF-train	0.851	0.869	0.835
	word-parser	0.679	0.781	0.600
	chunk-parser	0.786	0.769	0.803
DSTC3	CRF-train	0.606	0.567	0.649
	word-parser	0.552	0.685	0.462
	chunk-parser	0.817	0.786	0.851

Table 1: Evaluation of the zero-shot semantic parser in terms of F-score on the 1-best ASR hypothesis, Precision and Recall.

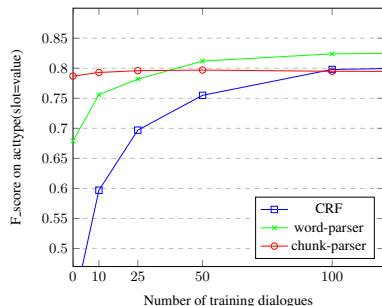


Figure 2: Impact of training data on the different semantic parsers on DSTC2.

a large amount of data (0.786 vs. 0.851 on DSTC2) the chunk-parser model perform better than CRF when just a small amount of data is available (0.817 vs. 0.606 on DSTC3).

Even if these results are not reported in Table 1, the chunk-parser has instantly a comparable performances<sup>4</sup> to those obtained by either rule-based system which has been used in the DSTC challenge (0.786 vs. 0.782 on DSTC2 and 0.817 vs. 0.824 on DSTC3) and also with the statistical model presented in [42] as SLU1 (0.786 vs. 0.803). So our proposed approach reaches close state-of-the-art performance without specific handcrafted rules (human expert cost) or training data (annotator cost).

Since the CRF model trained on a large amount of data obtains the best performance, we tried to show the impact of considering additional training data in our proposed approaches. For that, we propose to both add progressively some training annotated examples to the knowledge base and refine the considered feature matcher model used in our parser with the incoming training (in-context) data. In order to have a fair comparison we also consider a CRF model trained on the same training data.

Results presented in Figure 2 show that the chunk-parser is a good option in a situation where there is no or few training data. However, word-parser and CRF reach better performances after 50 (382 sentences) and 100 dialogues (782 sentences) respectively. Indeed, chunk features are more specific and less frequent in training data compared to words in feature matcher. These characteristics allow us to deal with the lack of in-context data at the starting point but decrease the robustness of the model when sufficient training data are available.

### 5.3. Generalisation

The major advantage of using a word2vec model is the integration of a continuous representation of words in the SLU decoding process. This characteristic endows the system with a natural generalisation capacity which can go as far as enabling to

<sup>4</sup>It is important that for both the rules and the Williams' parser the n-best lists of ASR hypothesis are considered as input instead of 1-best ASR in our experiment

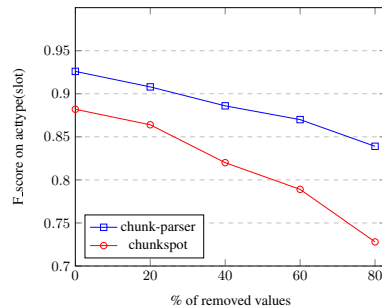


Figure 3: Generalisation ability of the zero-shot semantic parser approach on the DSTC2.

cover unknown words corresponding to unknown values for the defined ontology. For example, in the context of a restaurant search domain it is interesting for a dialogue system to detect some situations where a user talks about an unknown food type even if the latter is not in its original database for at least being able to propose an alternative accordingly.

In order to evaluate the generalisation ability of our system we removed from the DSTC2 knowledge base the OOCsFs corresponding to different percentages of the possible values of some specific slots. In this preliminary study, we chose to impact the *food*, *area* and *pricerange* slots. The model performances on manual transcriptions were evaluated in terms of F-score for *actype(slot)* only instead of *actype(slot=value)* in order to evaluate the high level concept elicitation. Thereby, we compare the chunk-parser performance with another configuration of the same parser, noted chunkspot. The latter is tuned to only tag chunks which reach a fix high similarity threshold (here a quite perfect matching - 0.94). Thus, as it is also able to slightly exploit the semantic space, this configuration can be assimilated to a more robust chunk spotting strategy. The results (presented in Figure 3) clearly show a slight drop in performance when the percentage of removed values grows with the full method. The difference between the two performances is equal to 0.044 at 0% and to 0.111 at 80%. It tends to confirm that the considered approach is more tolerant to data sparsity in  $K$  which can be useful for open domain dialogue system permitting a seamless evolution of the knowledge base populated by a growing database.

## 6. Conclusion and future works

In this paper an approach of zero-shot learning for SLU based on word embedding semantic space is proposed. We showed that such approach reach instantly performances which are comparable to those obtained with system trained on large data corpora or based on detailed handcrafted rules though. It can be developed at reduced cost. The proposed method appeared also more tolerant to unknown slot values and thus offers a valuable generalization mechanism which can be employed to on the fly domain extension. We also show that the quality of our proposed approach can be improved by additional training data.

However, the quality of both the knowledge-base  $K$  and the embedding  $F$  are important factors of the performance reached by our proposed method. Nevertheless, considering the proposed approach in the context of conditional random fields has the advantage of dealing with a model able to be tuned and adapted in an active learning strategy. This will be explored in coming works.

## 7. References

- [1] S. Hahn, M. Dinarelli, C. Raymond, F. Lefèvre, P. Lehnen, R. De Mori, A. Moschitti, H. Ney, and G. Riccardi, “Comparing stochastic approaches to spoken language understanding in multiple languages,” *TASLP*, vol. 19, no. 6, pp. 1569–1583, 2010.
- [2] F. Lefèvre, “Dynamic Bayesian networks and discriminative classifiers for multi-stage semantic interpretation,” in *ICASSP*, 2007.
- [3] A. Deoras and R. Sarikaya, “Deep belief network based semantic taggers for spoken language understanding,” in *INTERSPEECH*, 2013.
- [4] N. Camelin, B. Detienne, S. Huet, D. Quadri, and F. Lefèvre, “Unsupervised concept annotation using latent dirichlet allocation and segmental methods,” in *EMNLP Workshop on Unsupervised Learning in NLP*, 2011.
- [5] G. Tur, D. Hakkani-tur, D. Hillard, and A. Celikyilmaz, “Towards unsupervised spoken language understanding: Exploiting query click logs for slot filling,” in *INTERSPEECH*, 2011.
- [6] A. Lorenzo, L. Rojas-Barahona, and C. Cerisara, “Unsupervised structured semantic inference for spoken dialog reservation tasks,” in *SIGDIAL*, 2013.
- [7] A. Celikyilmaz, G. Tur, and D. Hakkani-Tur, “Leveraging web query logs to learn user intent via bayesian latent variable model,” in *ICML*, 2011.
- [8] D. Hakkani-Tur, L. Heck, and G. Tur, “Exploiting query click logs for utterance domain detection in spoken language understanding,” in *ICASSP*, 2011.
- [9] Y. Gao, L. Gu, and H. Kuo, “Portability challenges in developing interactive dialogue systems,” in *ICASSP*, 2005.
- [10] R. Sarikaya, “Rapid bootstrapping of statistical spoken dialogue systems,” *Speech Communication*, vol. 50, no. 7, pp. 580–593, 2008.
- [11] G. Tur, G. Rahim, and D. Hakkani-Tur, “Active labeling for spoken language understanding,” in *EUROSPEECH*, 2003.
- [12] G. Tur, D. Hakkani-Tur, and R. Schapire, “Combining active and semi-supervised learning for spoken language understanding,” *Speech Communication*, vol. 45, no. 2, pp. 171–186, 2005.
- [13] F. Lefèvre, F. Mairesse, and S. Young, “Cross-lingual spoken language understanding from unaligned data using discriminative classification models and machine translation,” in *INTERSPEECH*, 2010.
- [14] B. Jabaian, L. Besacier, and F. Lefèvre, “Comparison and Combination of Lightly Supervised Approaches for Language Portability of a Spoken Language Understanding System,” *TASLP*, vol. 21, no. 3, pp. 636–648, 2013.
- [15] F. Lefèvre, D. Mostefa, L. Besacier, Y. Esteve, M. Quignard, N. Camelin, B. Favre, B. Jabaian, and L. Rojas-Barahona, “Robustness and portability of spoken language understanding systems among languages and domains: the PORT-MEDIA project,” in *LREC*, 2012.
- [16] S. Zhu, L. Chen, K. Sun, D. Zheng, and K. Yu, “Semantic parser enhancement for dialogue domain extension with little data,” in *SLT*, 2014.
- [17] Y. Dauphin, G. Tur, D. Hakkani-Tur, and L. Heck, “Zero-shot learning and clustering for semantic utterance classification,” *arXiv preprint arXiv:1401.0509*, 2014.
- [18] E. Ferreira, B. Jabaian, and F. Lefèvre, “Online adaptive zero-shot learning spoken language understanding using word-embedding,” in *ICASSP*, 2015.
- [19] M. Henderson, B. Thomson, and J. Williams, “The second dialog state tracking challenge,” in *SIGDIAL*, 2014.
- [20] M. Henderson, B. Thomson and J. Williams, “The third dialog state tracking challenge,” in *SLT*, 2014.
- [21] C. Raymond and G. Riccardi, “Generative and discriminative algorithms for spoken language understanding,” in *INTERSPEECH*, 2007.
- [22] J. Lafferty, A. McCallum, and F. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” in *ICML*, 2001.
- [23] K. Yao, G. Zweig, M.-Y. Hwang, Y. Shi, and D. Yu, “Recurrent neural networks for language understanding,” *INTERSPEECH*, 2013.
- [24] G. Mesnil, X. He, L. Deng, and Y. Bengio, “Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding,” in *INTERSPEECH*, 2013.
- [25] K. Yao, B. Peng, G. Zweig, D. Yu, X. Li, and F. Gao, “Recurrent conditional random fields,” in *NIPS Deep Learning Workshop.*, 2013.
- [26] k. Yao, B. Peng, G. Zweig, D. Yu, X. Li, and F. Gao, “Recurrent conditional random field for language understanding,” in *ICASSP*, 2014.
- [27] M. Palatucci, D. Pomerleau, G. Hinton, and T. Mitchell, “Zero-shot learning with semantic output codes,” in *Advances in Neural Information Processing Systems 22*, 2009, pp. 1410–1418.
- [28] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.
- [29] J. Bian, B. Gao, and T. Liu, “Knowledge-powered deep learning for word embedding,” in *ECML*, 2014.
- [30] S. Bengio and G. Heigold, “Word embeddings for speech recognition,” in *INTERSPEECH*, 2014.
- [31] S. Clinchant and F. Perronnin, “Aggregating continuous word embeddings for information retrieval,” in *CVSC*, 2013.
- [32] W. Zou, R. Socher, D. Cer, and C. Manning, “Bilingual word embeddings for phrase-based machine translation,” in *EMNLP*, 2013.
- [33] T. Lavergne, J. M. Crego, A. Allauzen, and F. Yvon, “From n-gram-based to CRF-based translation models,” in *WSMT*, 2011.
- [34] B. Jabaian, F. Lefèvre, and L. Besacier, “A Unified Framework for Translation and Understanding Allowing Discriminative Joint Decoding for Multilingual Speech Semantic Interpretation,” *Computer Speech and Language*, 2014.
- [35] H. Larochelle, D. Erhan, and Y. Bengio, “Zero-data learning of new tasks,” in *Conference on Artificial Intelligence*, 2008.
- [36] L. Heck and D. Hakkani-Tur, “Exploiting the semantic web for unsupervised spoken language understanding,” in *SLT*, 2012.
- [37] T. Anastasakos and A. Deoras, “Task specific continuous word representations for mono and multi-lingual spoken language understanding,” in *ICASSP*, 2014.
- [38] T. Mikolov, W. Yih, and G. Zweig, “Linguistic regularities in continuous space word representations,” in *NAACL-HLT*, 2013.
- [39] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in Neural Information Processing Systems*, 2013.
- [40] T. Lavergne, O. Cappé, and F. Yvon, “Practical very large scale CRFs,” in *ACL*, 2010.
- [41] S. Huet and F. Lefèvre, “Unsupervised alignment for segmental-based language understanding,” in *Proceedings of the First Workshop on Unsupervised Learning in NLP*, 2011.
- [42] J. Williams, “Web-style ranking and slu combination for dialog state tracking,” in *SIGDIAL*, 2014.