ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/280086203

Online adaptative zero-shot learning spoken language understanding using word-embedding

Conference Paper · April 2015

DOI: 10.1109/ICASSP.2015.7178987

CITATIONS
8

READS
531

3 authors, including:

Emmanuel Ferreira
Université d´Avignon et des Pays du Vaucluse
12 PUBLICATIONS **75** CITATIONS

SEE PROFILE

Bassam Jabaian
Université d´Avignon et des Pays du Vaucluse
24 PUBLICATIONS **91** CITATIONS

SEE PROFILE

All content following this page was uploaded by Emmanuel Ferreira on 16 July 2015.

The user has requested enhancement of the downloaded file.

ONLINE ADAPTATIVE ZERO-SHOT LEARNING SPOKEN LANGUAGE UNDERSTANDING USING WORD-EMBEDDING

Emmanuel Ferreira, Bassam Jabaian and Fabrice Lefèvre [*]

CERI-LIA, University of Avignon, Avignon - France

ABSTRACT

Many recent competitive state-of-the-art solutions for understanding of speech data have in common to be probabilistic and to rely on machine learning algorithms to train their models from large amount of data. The difficulty remains in the cost and time of collecting and annotating such data, but also to update the existing models to new conditions, tasks and/or languages. In the present work an approach based on a zero-shot learning method using word embeddings for spoken language understanding is investigated. This approach requires no dedicated data. Large amounts of un-annotated and unstructured found data are used to learn a continuous space vector representation of words, based on neural network architectures. Only the ontological description of the target domain and the generic word embedding features are then required to derive the model used for decoding. In this paper, we extend this baseline with an online adaptative strategy allowing to refine progressively the initial model with only a light and adjustable supervision. We show that this proposition can significantly improve the performance of the spoken language understanding module on the second Dialog State Tracking Challenge (DSTC2) datasets.

Index Terms— Spoken language understanding, word embedding, zero-shot learning, out-of-domain training data, online adaptation.

1. INTRODUCTION

In dialogue systems, the Spoken Language Understanding (SLU) module extracts 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 trained on a large amount of data with various machine learning methods (see for instance [1, 2, 3] etc.). Despite their good performance, the difficulty remains in the cost and time of collecting and annotating such required data and also in the updating of the existing models to new conditions, tasks and/or languages. Therefore, some research works have focused on the use of lightly supervised [4, 5, 6], or unsupervised [7, 8, 9] training approaches to cope with the lack of annotated resources by either exploiting the semantic web for mining additional training data and enriching classification features or proposing unsupervised annotation process on a closed-domain corpus. Always with the objective of minimising the cost of data collection, some other works focused on porting a system across language and domain [10, 11, 12]. Active learning has been also widely studied as a way to reduce the time required for corpus annotation and verification in online settings [13, 14, 15].

In this work, a zero-shot learning technique is employed to deal with the SLU task. The proposed approach rests on a word embedding semantic modelisation which alleviates parts of the requirement in terms of annotated and in-context data by exploiting its internal generalisation properties [16]. 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 used at decoding time. Recently, a similar approach has been proposed in [17] for Semantic Utterance Classification (SUC). However, our proposition is different with respect to: how the semantic space is modelled, no in-context domain data required, and what is the task at hand, semantic annotation of a sentence (SLU) and not whole utterance classification (SUC). However, such approach is dependent on the quality of both the given ontological description and the word-embedding space. To address this limitation, we propose to add an online adaptative strategy, introducing a light supervision, in order to refine the initial knowledge base definition and to better exploit the considered embedding in an incremental fashion. This proposition joins recent works addressing the SLU adaptation issue. For instance, in [18] an instance-based approach for online adaptation of semantic models is presented, while [19] proposes a supervised approach for updating the SLU models with a limited supervision given by users calling the system. We show that our proposed online adaptative technique improves the performance reached by the zero-shot learning method in several configurations on the Dialog State Tracking (DSTC2) testbed [20].

In Section 2 of this paper we describe the considered SLU task. Section 3 presents the proposed adaptative zero-shot learning approach. Our experimental study is presented in Section 4 followed by some concluding remarks.

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

2. SPOKEN LANGUAGE UNDERSTANDING

The aim of the SLU module is to extract from a user utterance of n words, $W = w_1, w_2, ..., w_n$, a valid sequence of m concepts $C = c_1, c_2, ..., c_m$, where c_i is formally described as a slot-value pair such as *food=Italian* or *destination=Boston*. However, in this paper, the semantics follows the standards defined during the challenges embodied in the DSTC2 and DSTC3 corpora [20] wherein the extracted sequence of concepts is expressed as a sequence of Dialogue Acts (DAs) of the form `acttype(slot=value)`.

Thus, the considered SLU task is a sequential tagging problem where possible tags are all task-specific `acttype( slot=value)` combinations based on a pre-defined inventory of acttypes, slots and associated values.

The `acttype` are task-independent and can be divided into 4 groups: information providing (`inform`), query (`request, reqalts, 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)`".

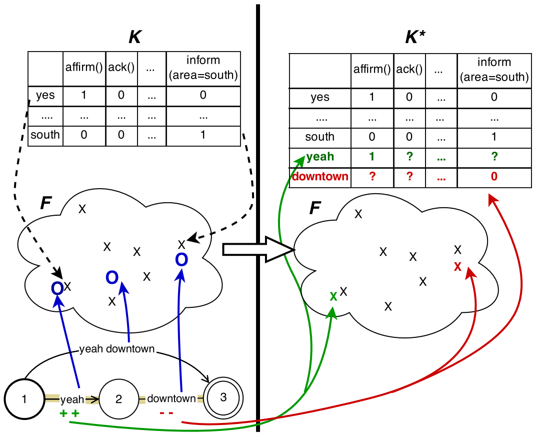
3. TOWARDS ONLINE ADAPTATIVE ZERO-SHOT LEARNING

The zero-shot learning, as proposed in [21], corresponds to a learning scheme where possible values for a given class include cases that have been omitted from the training examples. In this study, we focus on the problem of predicting the semantic tag sequence of a user query without having seen any example of in-domain user utterances and thus in-context semantic tags. This is done by defining a semantic knowledge mapping between in-domain and general knowledge data in order to extrapolate these tags.

3.1. Zero-learning for SLU

The initial model, depicted in Figure 1, makes use of three main components. The first one is a semantic feature space F based on word embedding learnt with neural network algorithms [16, 22]. This representation offers a continuous representation of word. Indeed, several researches pinpointed the interest of exploiting some regularities between syntactic/semantic features of words and their corresponding embedding for different NLP tasks (e.g. [23]). The objective is to define a metric space encoding the semantic properties of all possible tags.

The second component is the semantic knowledge base K that corresponds to a domain-specific assignment table (as shown in Fig. 1) where each row represents the assignments to each possible semantic tags (columns) for a specific d-dimensional vector in F. Such kind of vectors are obtained by projecting into F some lexical items, called chunks hereafter as they may include compounds of words extracted from the ontological description of the domain. For example, "what food is served?" for `request(food)`, or "french food" for `inform(food=french)`. These chunks can be easily obtained by an automatic process over the ontology and the backend database combined with a few manually defined dialogue acts examples (as `acttypes` are not present in the ontology).



	affirm()	ack()	...	inform (area=south)
yes	1	0	...	0
....	...	...	...	...
south	0	0	...	1

	affirm()	ack()	...	inform (area=south)
yes	1	0	...	0
....	...	...	...	...
south	0	0	...	1
yeah	1	?	...	?
downtown	?	?	...	0

Fig. 1. Illustration of the zero-shot learning SLU parsing and adaptation

Thus, in Fig. 1, K's rows and columns are labelled with chunks and semantic tags for convenience. Basically the value in each cell $c_{i,j}$, denoted as the *assignment value* hereafter, indicates if any assignment exists between the i^{th} vector in the semantic space and the j^{th} semantic tag. The assignment values are initiated with binary values. Notice that we do not constrain the current representation to one-to-one mapping between chunks and tags, so several assignments could be not null in a single row. For instance, the chunk "Paris" could be associated to both `inform(location=Paris)` and `inform(name=Paris)` if a venue happens to be named Paris.

The third component of our proposed baseline model is the SLU parser. It extracts a scored graph of semantic tag sequence hypotheses from any novel user utterance. All contiguous word sequences (chunks) are considered in the parsing algorithm. For example if the user says "yeah downtown", as in Fig. 1, 3 different chunks are considered: "yeah", "downtown" and "yeah downtown". These chunks are mapped to the feature space F with the same method used to map K's chunks into F. The resulting vectors (blue circles in Fig. 1) are then compared in terms of similarity (e.g. cosine similarity) to the known chunk vectors (black crosses in Fig. 1). Then, a dot product between the similarity vector and K matrix is computed and a k-nearest neighbors classifier is employed to attribute to each chunk the ordered semantic hypotheses (one for each possible tags). These latter are then employed to construct a finite-state transducer where chunks and their corresponding semantic hypotheses are the edge labels, weighted by the distances (inverses of the observed similarities). A final rescaling process allows to balance the influence of chunk lengths. The best semantic utterance hypothesis is obtained by a shortest-path decoding of the finite-state machine (highlighted path in Fig. 1).

3.2. Online adaptation

Based on the baseline SLU parser, an online adaptation process is defined with which we intend to refine incrementally the assignment values in K according to some feedbacks gathered during live usage of the system. Indeed, even if the performance being generally measured only on a bag of tags, a direct word (or chunk) to tag association can be retrieved from the finite-state transducer produced by the SLU parser and thus can be employed to adapt the model with new information.

In order to minimize the supervision effort, an adaptation scheme where the supervision is limited to binary feedbacks (validation-refutation) is envisioned. This technique, involving no manual corrections to assign the true labels, can be easily integrated in an existing dialogue setup and allows to master the ratio cost/improvement through the quota of feedbacks asked to the user at each turn. Defining an optimal strategy with respect to this ratio will be addressed in further works. In the present work, only improvement is targeted and a user giving all feedbacks possible at each turn is simulated. These feedbacks are then employed to update K to K^*.

An example of that process is given in Fig. 1. It illustrates a case where the true semantic labels of the user utterance are misrecognized by the parsing strategy: the utterance "yeah downtown" is tagged as `affirm(), inform(area=south)` instead of `affirm(), inform(area=centre)`.

The user feedbacks populate a set of m tuples $U = ((c_k, T_k, f_k))_{1 \leq k \leq m}$, where ($c_k$, T_k) is the chunk/tag pair proposed to the user and f_k is her feedback (1 positive, 0 negative). Given K and U after each interaction Algorithm 1 (partially illustrated in Fig.1) is used to update K to K^*. Each cell (i, j) in K corresponds to a chunk (row)/tag (column) pair and encloses 4-values: $p_{i,j}$ and $n_{i,j}$ represent respectively the number of observed positive and negative feedbacks up to now, $knn_{i,j}$ is the value obtained by computing an element-wise addition of the k nearest neighbors rows (rescaled via a dot product of the normalised similarity of these rows to the i^{th} chunk - Alg. 1.16) and $c_{i,j}$ is the assignment value, exploited by the parsing algorithm presented above. The algorithm shows how K is extended with new rows and how every $c_{i,j}$ is updated. Basically, first a new row is added to K each time an unseen chunk c_k is found in U (see Alg. 1.4-6). Then we update all the feedback counts based on U (see Alg. 1.8-9). For that purpose two scaling factors α_p and α_n allow to scale the importance of the new information, and can be set to different values to distinguish safe data collections with trustful users from normal field online adaptation. For the initial set of chunk/tag pairs, $p_{i,j}$s are initiated with a prior p_0. So in the general case the assignment value is obtained as a ratio of the positive/negative feedbacks associated to it (see Alg. 1.12).

Algorithm 1 Knowledge base update

```
 1: Given: $K$ and $U$ Output: $K^*$
 2: $K^* \leftarrow K$
 3: for all $(c, T, f) \in U$ do
 4:     if $c \notin K^*$ then
 5:         append new row for $c$ in $K^*$ with default cells
 6:         $m_{last} = 1$
 7:     $i \leftarrow c$-row id,  $j \leftarrow T$-colunm id
 8:     $p_{i,j} \leftarrow p_{i,j} + f \times \alpha_p$
 9:     $n_{i,j} \leftarrow n_{i,j} + (1 - f) \times \alpha_n$
10:     if $p_{i,j} + n_{i,j} > 0$ then
11:         $old_c \leftarrow c_{i,j}$
12:         $c_{i,j} \leftarrow \frac{p_{i,j}}{p_{i,j}+n_{i,j}}$
13:         if $c_{i,j} - old_c < 0$ then  $m_i \leftarrow 1$
14:     else $c_{i,j} \leftarrow 0$
15: for all $c_{i,j} \in K^*$ do
16:     compute $knn_{i,j}$
17: for all $c_{i,j} \in K^*$ do
18:     if $p_{i,j} + n_{i,j} = 0$ and $m_i = 1$ then  $c_{i,j} \leftarrow knn_{i,j}$
```

For each row a modification flag m_i is used to detect if a prior knowledge (positive assignment) is challenged by new evidences (measured by a decrease in $c_{i,j}$, see Alg. 1.13). In that case the other possible assignments for this chunk (other cells in the row) have their assignment values set to the knn value, instead of 0, so that new associations can be tested and proposed for evaluation to the user.

4. EXPERIMENTS AND RESULTS

All experiments presented in the paper are based on the DSTC2 datasets [20] covering the domain of restaurant search. Event if this research challenge focused on tracking the user's goal all along the dialogue, here we only consider the SLU task. Thus, we exploit the fully annotated data (e.g. transcriptions, dialogue-act semantics) as train and test sets to evaluate our adaptative zero-shot semantic decoding approach on realistic dialogue settings. In our experiment, we evaluate the approach on the given 10-best ASR of the challenge test set (9890 user utterances). A subset of transcriptions from the DSTC2 training set (up to 1472 transcribed user utterances) is also exploited to simulate the online adaptation presented in 3.2.

To define the semantic space, the word2vec [16] word-embedding model is considered. A 300-dimensional model was trained on a large amount of wide coverage and freely available English corpora[1] with the Skip-gram algorithm (with a 10-word window). The resulting model is expected to exhibit some linguistic regularities as those shown in [24] as well as a linear structure that makes it possible to meaningfully combine the words by an element-wise addition of their word embeddings [25]. So the latter technique is employed to directly map word chunks to their corresponding word2vec representation sees as the sum of individual word representations. Due to fact that word2vec behaves nicely with cosine similarity in the literature [16, 24], this metric is prefered in a k-nearest neighbors classifier for the chunk prediction and extension (in the following experiments $k = 1$ for parsing and 20 for knn cell values). We employed the shortest-path algorithm on the semantic graph with the cosine distance (1 - cosine similarity) metric (see Section 3.1).

The task-dependent knowledge base used in the experiments is derived from the challenge's ontology, as well as from a generic dialogue information. The semantics of the DSTC2 task is represented by 16 different act types, 8 slots and 215 values. The lexical forms (53) used to model act types were manually written (for example "say again" for the `repeat` act). In the considered ontology, slots and values have already lexicalised names (e.g. "address", "french", etc.). Overall, 4160 automatically generated chunks are considered and assigned to 663 different semantic tags.

In order to compare the online adaptative capacity of the zero-shot learning algorithm for SLU, two baselines are considered: a rule-based system used in the DSTC challenge (noted Rules-b) and one learnt on the DSTC2 training data (referred to as SLU1 in [26] and noted Learnt-b hereafter). Three different Zero-Shot Semantic Parser (ZSSP) configurations are evaluated to contrast the influence of both the semantic space F and the knowledge base K.

First, a classic ZSSP uses a qualitative lightly handcrafted K and a robust word2vec semantic space (as described above); second, ZSSP.$\tilde{F}$ uses a word2vec space limited to a 50-dimensional model learnt on a small in-domain training set (DSTC2 training set with user and system utterances); finally ZSSP.$\tilde{K}$ uses a downgraded (cheaper) version of K where 10% among the manually written forms were pruned.

In order to determine the impact of the online adaptative scheme described in Section 3.2, transcribed utterances from the DSTC2 train set are used to simulate online adaptation (avoiding noise due to ASR errors) and the test set (10-best ASR user inputs) for evaluation. For adaptation, the user feedbacks are simulated by comparing the top-hypothesis of the current model to the reference semantic label in the DSTC2 annotations for each proposed chunk. All semantic tags of the top SLU hypothesis present in the true semantic sequence are considered as positive and all others as negative. K is updated at the end of each turn (with $\alpha_p = \alpha_n = 1$).

Results in Figure 2 show the evolution of the F-score according to the number of adaptation dialogues. Before adaptation ZSSP (0.794) and ZSSP.$\tilde{K}$ (0.775) reach close to the Rules-b performances (0.782) without specific handcrafted rules (human expert cost) or training data (annotator cost). A semantic space learnt on a little amount of non-general data can significantly impact the initial performance as shown with ZSSP.$\tilde{F}$ (0.684) due to both out-of-vocabulary words and bad generalisation properties of this semantic space.

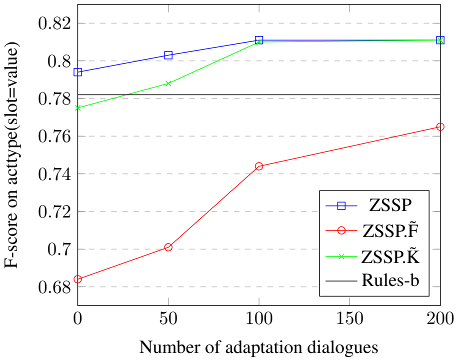


Fig. 2. Refined performance in terms of F-score on DSTC2 test set of various configuration of the online adaptation method according to the number of dialogues.

Nevertheless, in all ZSSP configurations, the performance grows jointly with the number of adaptation dialogues. Indeed, both ZSSP and ZSSP.$\tilde{K}$ configuration obtain performance significantly better than the two baselines after only 100 dialogues: 0.811 for the two ZSSP methods vs. 0.782 for Rules-b and 0.803 for Learnt-b (not showed in Fig. 2). Moreover, even the gap between ZSSP.$\tilde{F}$ and Rules-b is clearly reduced all along the online adaptation (from 0.098 to 0.017 after 200 dialogues). This particular point demonstrates that the proposed method can also deal with unfitted semantic space. These overall results flag the benefit of the proposed online adaptation method to cope with both the limitations of the initial K coverage and F robustness.

5. CONCLUSION AND FUTURE WORKS

In this paper a method for zero-shot learning SLU is proposed and tested. It is shown that such method reaches state-of-the-art performance on the DSTC2 task. In particular the extension of the method to an online adaptation scheme has been proved to be efficient and to provide a practical way to alleviate some of the limitations inherent to a zero-sot learning approach based on a word embedding semantic space such as the initial quality of both the semantic space and the knowledge base. The supervision effort still remains very low since the user is just asked to confirm some hypotheses made by the system but never to explicitly correct any error. However, comparison with other active learning techniques and generalisation of the approach in a reinforcement learning framework are in progress and its integration in a live dialogue system should be presented soon.

[1] enwik9, One Billion Word Language Modelling Benchmark, the Brown corpus, English GigaWord from 1 to 5

6. 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," *IEEE 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] A. Celikyilmaz, G. Tur, and D. Hakkani-Tur, "Leveraging web query logs to learn user intent via bayesian latent variable model," in *ICML*, 2011.

[5] D. Hakkani-Tur, L. Heck, and G. Tur, "Exploiting query click logs for utterance domain detection in spoken language understanding," in *ICASSP*, 2011.

[6] L. Heck and D. Hakkani-Tur, "Exploiting the semantic web for unsupervised spoken language understanding," in *SLT*, 2012.

[7] 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.

[8] 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.

[9] A. Lorenzo, L. Rojas-Barahona, and C. Cerisara, "Unsupervised structured semantic inference for spoken dialog reservation tasks," in *SIGDIAL*, 2013.

[10] R. Sarikaya, "Rapid bootstrapping of statistical spoken dialogue systems," *Speech Communication*, vol. 50, no. 7, pp. 580–593, 2008.

[11] 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.

[12] B. Jabaian, L. Besacier, and F. Lefèvre, "Comparison and Combination of Lightly Supervised Approaches for Language Portability of a Spoken Language Understanding System," *IEEE TASLP*, vol. 21, no. 3, pp. 636–648, 2013.

[13] G. Tur, G. Rahim, and D. Hakkani-Tur, "Active labeling for spoken language understanding.," in *EUROSPEECH*, 2003.

[14] P. Gotab, F. Béchet, and G. Damnati, "Active learning for rule-based and corpus-based spoken language understanding models," in *ASRU*, 2009.

[15] F García, L. Hurtado, E. Sanchis, and E. Segarra, "An active learning approach for statistical spoken language understanding," in *CIARP*, 2011.

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[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] A. Bayer and G. Riccardi, "On-line adaptation of semantic models for spoken language understanding," in *ASRU*, 2013.

[19] P. Gotab, G. Damnati, F. Béchet, and L. Delphin-Poulat, "Online slu model adaptation with a partial oracle," in *INTERSPEECH*, 2010.

[20] M. Henderson, B. Thomson, and J. Williams, "The second dialog state tracking challenge," in *SIGDIAL*, 2014.

[21] Mark Palatucci, Dean Pomerleau, Geoffrey E. Hinton, and Tom M. Mitchell, "Zero-shot learning with semantic output codes," in *Advances in Neural Information Processing Systems 22*, pp. 1410–1418. 2009.

[22] J. Bian, B. Gao, and T. Liu, "Knowledge-powered deep learning for word embedding," in *ECML*, 2014.

[23] S. Clinchant and F. Perronnin, "Aggregating continuous word embeddings for information retrieval," in *Workshop on Continuous Vector Space Models and their Compositionality*, August 2013.

[24] T. Mikolov, W. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in *NAACL-HLT*, 2013.

[25] 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.

[26] Jason D Williams, "Web-style ranking and slu combination for dialog state tracking," in *Meeting of the Special Interest Group on Discourse and Dialogue*, 2014.

View publication stats