Attention-based Semantic Priming for Slot-filling

Jiewen Wu[1,2], **Rafael E. Banchs**[2], **Luis Fernando D'Haro**[2],
Pavitra Krishnaswamy[2], **and Nancy Chen**[2]

[1]A*STAR AI Initiative, Singapore
[2]Institute for Infocomm Research, A*STAR, Singapore
{wujw, rembanchs, luisdhe, pavitrak, nfychen}@i2r.a-star.edu.sg

Abstract

The problem of sequence labelling in language understanding would benefit from approaches inspired by semantic priming phenomena. We propose that an attention-based RNN architecture can be used to simulate semantic priming for sequence labelling. Specifically, we employ pre-trained word embeddings to characterize the semantic relationship between utterances and labels. We validate the approach using varying sizes of the ATIS and MEDIA datasets, and show up to 1.4-1.9% improvement in F1 score. The developed framework can enable more explainable and generalizable spoken language understanding systems.

1 Introduction

Priming (Waltz and Pollack, 1985) is a cognitive mechanism in which a primary stimulus (i.e. the prime) influences the response to a subsequent stimulus (i.e. the target) in an implicit and intuitive manner. In the case of *semantic priming*, both the prime and the target typically belong to the same semantic category. Semantic priming can be explained in terms of induced activation in associative neural networks (McClelland and Rogers, 2003). Further, there is empirical evidence to suggest that the processing of words in natural language is influenced by preceding words that are semantically related (Foss, 1982). Therefore, semantic priming approaches would enable improvements in sequence labelling.

Previous studies have leveraged contextual information in utterance sequences (Mesnil et al., 2015) and dependencies between labels (Ma and Hovy, 2016) to improve performance in sequence labelling tasks. However, there is limited work to use contextual information in utterances to inform inference of the subsequent labels through semantic priming. For instance, "*I'd like to book ...*" not only suggests the next word(s), e.g., flight, but also the label of the next word(s), e.g., *services*. We posit that systems employing this mode of cross-linked semantic priming could enhance performance in a variety of sequence labelling tasks.

In this work, we hypothesize that semantic priming in human cognition can be simulated by means of an attention mechanism that uses word context to enhance the discriminating power of sequence labelling models. We propose and explore the use of attention (Bahdanau et al., 2014) in a deep learning architecture to simulate the semantic priming mechanism. We apply this concept to slot filling, an example of sequence labelling in spoken language understanding, which aims to label the utterance sequences with a set of begin/in/out (BIO) tags. Specifically, we use pre-trained word embeddings to characterise not only the context of words, but also the semantic relationship between words in utterances and words in labels.

Overall, we develop a semantic priming based approach for the task of slot-filling to associate utterances and label sequences. Our contributions are as follows: (1) We propose an approach that applies semantic priming to sequence labelling. To capture semantic associations between utterance words and label words, we use three different strategies for deriving label embeddings from pre-trained embeddings. (2) We implemented the approach in an LSTM-based architecture and validate the efficacy of the approach.

In Section 2 we review related work. Section 3 elaborates the proposed approach. An empirical evaluation is provided in Section 4. Finally, Section 5 concludes the paper.

2 Related Work

Our proposed method draws on the attention mechanism, which has shown to be effective for

Proceedings of the Seventh Named Entities Workshop, pages 22–26
Melbourne, Australia, July 20, 2018. ©2018 Association for Computational Linguistics

sequence-based NLP tasks, particularly, machine translation (Bahdanau et al., 2014; Luong et al., 2015). Since attention allows the neural networks to dynamically attend to important features in the inputs, it is a suitable mechanism to achieve the objective of semantic priming between utterances and labels. Conditional random field (CRF) has been used together with RNNs, sometimes also including CNNs, to improve accuracy (Mesnil et al., 2013, 2015; Ma and Hovy, 2016; Reimers and Gurevych, 2017b). Dinarelli et al. (2017) proposes to learn label embedding for improving tagging accuracy, while our label embedding is computed directly from pre-trained word embeddings. Furthermore, our approach does not require shifted label sequences as input.

To use external knowledge, previous studies consider graph or entity embedding (Huang et al., 2017; Chen et al., 2016; Yang and Mitchell, 2017), together with other contextual information, such as dependency graph (Huang et al., 2017) or sentence structures (Chen et al., 2016). Specifically, Yang and Mitchell (2017) extends LSTM with graph embedding to learn concepts from knowledge bases and integrate the concept embedding into the state vectors of words. In contrast, our approach does not learn or parse sentences to get extra contextual information, which is suitable for languages lacking well trained parsers. Moreover, context integration is achieved without fine-tuning the underlying RNN structure yet rather through the attention mechanism.

3 Semantic Priming

Figure 1 depicts an LSTM-based neural network architecture for semantic priming. Given an utterance, a priming matrix is computed to connect the labels to input features generated by a bi-directional LSTM. The priming effects are then used for prediction.

3.1 Computing Priming Matrix

This section considers three different strategies of the proposed attention-based semantic priming mechanism. In all the three cases the input words are compared to proxies of the semantic categories over word vectors.

Let m denote the number of labels. An utterance of length n is represented by the matrix $X : n \times k$, where k is the dimension of pre-trained word vectors. Given a word vector x_j, semantic priming is achieved by comparing x_j with a label embedding matrix $L : m' \times k$, with m' unique concepts, each encoded in k dimensions. In addition, let $E_{l_i, 1 \le i \le m'}$ denote the set of embedded words tagged with the label l_i in the dataset. Note that the corresponding embedding of l_i is L_i. Below are the definitions of three different strategies to compute the label embeddings L.

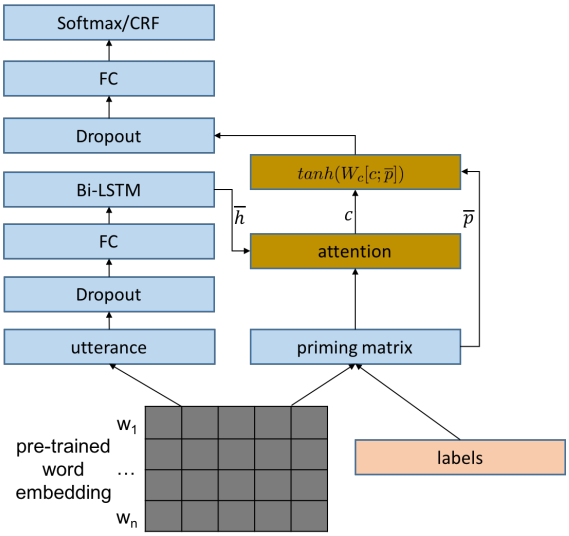

Figure 1: Proposed topology for priming. FC denotes a fully connected layer.

- Priming using Instance Centroid (**PIC**): L is defined to be $m \times k$ and $L_i = mean(E_{l_i})$. Intuitively, the proxy of the concept, L_i, is the centroid (mean vector) of the cluster of all known instance words in the concept.
- Priming using Instance Neighbor (**PIN**): L is defined to be $m \times k$ and

$$L_i = \underset{\forall e \in E_{l_i}}{\operatorname{argmin}}(1 - cos(x_j, e))$$

  In this case, the proxy of the concept is the nearest instance having the same label as x_j.
- Priming using Concepts (**PC**): L is defined to be $m' \times k$, m' is pre-specified, and $L_i = \mathbf{c}_i$, where c_i is a manually selected concept from l_i. The embedding representation, $\mathbf{c}_i$, is of dimension k as it is either the word vector per se of a single concept label or the mean vector of a set of such word vectors.

While **PIN** is a straightforward simulation of the semantic priming mechanism between a prime and its potential targets in different classes, **PIC** and **PC** are variants of a categorization mechanism referred to as the Basic Level (Rosch et al., 1976), in which the targets are intermediate, dominant concepts that represent the category.

	ATIS	MEDIA
# utterances in train	3982	12908
# utterances in dev	995	1259
# utterances in test	893	3005
# labels	127	138
vocab. size†	572	1671
max utterance length	46	192

Table 1: Statistics of datasets. †The vocabulary is a mix of words and entities.

Once L is computed, the priming matrix is computed by the cosine similarity, or the induced distance, between the word embedding of the utterance and L, i.e., $\overline{p} = cos(X, L)$.

3.2 Attention to Semantic Priming

In Figure 1, the hidden states, $\overline{h}$, of the bi-directional LSTM are considered to be the source, while the priming matrix $\overline{p}$ is analogous to the target. Following (Luong et al., 2015), we define the alignment scoring function to be $s(\overline{p}, h) = \overline{p} W_a h$ and compute the final output as follows:

$$\alpha = \frac{exp(s(\overline{p}, h))}{\sum_{i=1}^{n} exp(s(\overline{p}, h_i))}$$
$$c = \sum_{h} \alpha h$$
$$t = tanh(W_c[c; \overline{p}])$$

4 Experiments

To validate the efficacy of the architecture in Figure 1, an empirical evaluation was performed and implemented in Keras[1]. This section elaborates the experimental setup and presents our results.

4.1 Datasets

Two datasets on spoken dialogues were used in the experiments, namely, the Air Travel Information System (ATIS) task (Dahl et al., 1994) and MEDIA, French dialogues collected by ELDA (Bonneau-Maynard et al., 2005). The statistics of the two datasets is given in Table 1. For MEDIA, using entities significantly impacts the performance. Thus entities are used together with words in utterances, as implied by the size of vocabulary in Table 1. Since bi-directional LSTM is used in the architecture in Figure 1, no context word windows (Mesnil et al., 2015) were used as additional inputs in the datasets. The pre-trained word embedding sources for the two datasets are GloVe (English) (Pennington et al., 2014) and fastText (French) (Bojanowski et al., 2016), respectively. In particular, we found that there are about 100 words missing in the fastText French word embedding. Some of the words, however, are due to original tokenization in MEDIA.

4.2 Setup and Hyperparameters

To facilitate mini-batching for training, the utterances were padded to the maximum utterance length. For all experiments, we use one set of fixed hyperparameters to enable meaingful comparison. The dimension of word embedding is 300 for both GloVe and fastText. Following the recommendations in (Reimers and Gurevych, 2017a), all dropout layers have a rate of 0.5, and LSTM has an additional recurrent dropout of 0.5 between recurrent units. During learning phase, a mini-batch size of 18 and an initial learning rate of 0.004 was used with the Adam optimizer to minimize the cross-entropy loss. The learning rate was reduced by 50% after no improvement in three epochs.

As semantic priming provides connections between words and labels through the use of the same pre-training embedding, it will enable more robust performance even when the datasets are small. To validate this, we investigated the effects of semantic priming in cases where the datasets are reduced. Note that both ATIS and MEDIA have many short utterances; in particular, MEDIA has over 4000 utterances consisting of a single word. For reduction, we rank vocabulary by word frequency in the training and development sets and choose utterances containing the words until 100% of vocabulary is covered.

4.3 Results

In this section the conlleval-F1[2] scores are reported. The experiments were run on a NVIDIA DGX1 station (Tesla V100 and 16GB memory), and the F1 scores are the average of that in the first 30 epochs in three independent runs.

The results shown are for baseline with trainable embedding (**BE**), baseline with pre-trained embedding (**BP**), and the strategies defined in Section 3.1, i.e., **PIC**, **PIN** and **PC**. For **PC**, the concepts are the keywords that have occurred in the labels. Example concepts include *airline* in ATIS and *chambre* in MEDIA. A total of 30 and 53 concepts are extracted for **PC** in ATIS and MEDIA, respectively.

[1] https://keras.io/

[2] http://www.cnts.ua.ac.be/conll2000/chunking/output.html

Although **BE** yields much higher F1, we compare the proposed approach with the baseline approach, **BP**, where F1 is computed using pre-trained embedding. This is because all strategies, except for **BE**, are based on pre-trained word embedding. We also compare the results in the MEDIA dataset with and without CRF. Since CRF in ATIS was shown to lead to no improvement (Dinarelli et al., 2017), so, no CRF layer was applied to ATIS in the experiments.

	ATIS	MEDIA	
BP	94.22	**72.66**	79.46†
PIC	94.23	69.37	80.49†
PIN	94.41	69.79	**80.56**†
PC	**94.51**	72.55	78.35†
BE	94.75	82.16	86.38†

Table 2: F1 of the two datasets. †CRF used.

Table 2 shows the F1 computed over the full datasets. In ATIS, although no significant conclusions can be drawn, all strategies, in particular, **PC**, outperform the baseline **BP**. Note that, when CRF, instead of SOFTMAX, is used in MEDIA, there is an increase of 4% for **BE**, 7% for **BP**, and 10% for **PIC**/**PIN**. For MEDIA, F1 has a considerable drop when pre-trained word embedding is used instead of trainable embedding. When SOFTMAX is used, none of the strategies outperformed the baselines **BP** or **BE**. In contrast, once CRF is used both **PIC** and **PIN** gained over 1% increase compared with **BP**.

	ATIS_{100}	MEDIA_{100}	
BP	85.39	67.64	76.95†
PIC	**87.25**	66.84	76.81†
PIN	86.31	**68.25**	**78.34**†
PC	87.01	67.37	77.95†
BE	86.04	78.81	83.77†

Table 3: F1 of the reduced datasets. †CRF used. 100% of the vocabulary in datasets are retained.

Table 3 describes the results over reduced datasets that cover the full (100%) vocabulary in the datasets. ATIS_{100} has a total of 583 utterances for training/development, while MEDIA_{100} has 1717 for training/development. Note that reduction was *not* performed to test datasets, i.e., full test sets were used. For both ATIS and MEDIA, **PIN** shows consistent performance gain (+1%) over the pre-trained baseline approach (**BP**).

	ATIS_{70}	MEDIA_{70}	
BP	83.21	65.37	76.34†
PIC	83.23	**66.44**	75.2†
PIN	82.65	66.09	**77.12**†
PC	**83.4**	65.75	75.4†
BE	81.62	76.3	80.3†

Table 4: F1 of the reduced datasets. †CRF used. 70% of the vocabulary in datasets are retained.

Table 4 describes the results over further reduced datasets, i.e., these two reduced datasets covers only 70%[3] of the whole vocabulary, containing 348 and 1216 utterances (train/dev) for ATIS and MEDIA, respectively. As shown in Table 4, **PC** was the best strategy for ATIS while **PIN** consistently outperformed the baseline **BP** in MEDIA.

Overall, we have seen performance gains when priming is used over the original and reduced datasets, compared to the pre-trained baseline approach **BP**. In particular, we recommend **PIN** over the other strategies as it is less computational expensive compared with **PIC** while it seems to provide more consistent improvement over **BP** than other strategies.

5 Conclusions and Future Work

We have demonstrated an approach to leverage semantic priming for natural language understanding tasks. The approach employs pre-trained embeddings to prime label concepts based on utterance words. Our experimental results suggest improvements over baselines are feasible. However, we note that the coverage of the dataset vocabulary in the pre-trained word embedding may limit performance improvements. For example, the missing words in the pre-trained French word embedding adversely affected the F1 scores for MEDIA. The approach can be easily adapted to a variety of different network architectures (e.g., (Dinarelli et al., 2017)) and word embeddings (e.g., (Reimers and Gurevych, 2017a)). Future studies will focus on how to choose a good set of concepts for the PC priming strategy. It will also be fruitful to understand how to explain the sequence labelling outputs using attention mechanisms.

[3]70% allows for a considerable reduction of the full vocabulary yet not resulting in too small datasets.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Helene Bonneau-Maynard, Sophie Rosset, Christelle Ayache, A Kuhn, and Djamel Mostefa. 2005. Semantic annotation of the french media dialog corpus.

Yun-Nung Chen, Dilek Z. Hakkani-Tür, Gökhan Tür, Asli Çelikyilmaz, Jianfeng Gao, and Li Deng. 2016. Knowledge as a teacher: Knowledge-guided structural attention networks. *CoRR*, abs/1609.03286.

Deborah A. Dahl, Madeleine Bates, Michael Brown, William Fisher, Kate Hunicke-Smith, David Pallett, Christine Pao, Alexander Rudnicky, and Elizabeth Shriberg. 1994. Expanding the scope of the atis task: The atis-3 corpus. In *Proceedings of the Workshop on Human Language Technology*, HLT '94, pages 43–48, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marco Dinarelli, Vedran Vukotic, and Christian Raymond. 2017. Label-dependency coding in simple recurrent networks for spoken language understanding. In *INTERSPEECH*.

Donald Foss. 1982. A discourse on semantic priming. *Cognitive Psychology*, 14:590?607.

Lifu Huang, Avirup Sil, Heng Ji, and Radu Florian. 2017. Improving slot filling performance with attentive neural networks on dependency structures. In *EMNLP*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *EMNLP*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

James L. McClelland and Timothy T. Rogers. 2003. The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, 4:310–322.

Grégoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Hakkani-Tur, Xiaodong He, Larry Heck, Gokhan Tur, Dong Yu, and Geoffrey Zweig. 2015. Using recurrent neural networks for slot filling in spoken language understanding. *Trans. Audio, Speech and Lang. Proc.*, 23(3):530–539.

Grégoire Mesnil, Xiaodong He, Li Deng, and Yoshua Bengio. 2013. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In *INTERSPEECH*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Nils Reimers and Iryna Gurevych. 2017a. Optimal hyperparameters for deep lstm-networks for sequence labeling tasks. *CoRR*, abs/1707.06799.

Nils Reimers and Iryna Gurevych. 2017b. Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 338–348, Copenhagen, Denmark.

Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive Psychology*, 8(3):382 – 439.

David L. Waltz and Jordan B. Pollack. 1985. Massively parallel parsing: A strongly interactive model of natural language interpretation. *Cognitive Science*, 9(1):51 – 74.

Bishan Yang and Tom M. Mitchell. 2017. Leveraging knowledge bases in lstms for improving machine reading. In *ACL*.