CACHE NEURAL NETWORK LANGUAGE MODELS BASED ON LONG-DISTANCE DEPENDENCIES FOR A SPOKEN DIALOG SYSTEM

F. Zamora-Martínez[⋆†] *S. España-Boquera*[†] *M.J. Castro-Bleda*[†] *R. De-Mori*[‡]

[⋆] ESET, Universidad CEU-Cadenal Herrera, Valencia, Spain.
[†] DSIC, Universitat Politècnica de València, Valencia, Spain
[‡] LIA University of Avignon, Avignon, France

ABSTRACT

The integration of a cache memory into a connectionist language model is proposed in this paper. The model captures long term dependencies of both words and concepts and is particularly useful for Spoken Language Understanding tasks. Experiments conducted on a human-machine telephone dialog corpus are reported, and an increase in performance is observed when features of previous turns are taken into account for predicting the concepts expressed in a user turn. In terms of Concept Error Rate we obtained a statistically significant improvement of 3.2 points over our baseline (10% relative improvement) on the French Media corpus.

1. INTRODUCTION

The purpose of Language Models (LMs) in Automatic Speech Recognition (ASR) systems is to compute the probability of a word w given its history h, which is defined as the sequence of words uttered before w. If the process of ASR decoding is performed on a human/human conversation or a human/computer dialog, the history of a word may be very long and the estimation of the probability $P(w|h)$ may be very difficult due to the immense variety of possible histories. A popular solution is to approximate histories by the $n-1$ words preceding w in word n-grams. Even in this case, the estimation accuracy is affected by data sparseness. Motivated by the above considerations, a solution is proposed based on continuous space LMs and effective approximations of word histories made of summaries composed of a limited number of semantic constituents useful for word prediction. History summaries are made of discourse features. In [1] intentions and preferentially retained information are considered to model the attentional state of a conversation. A cache model is proposed for temporarily storing this information. Inspired by these ideas and by a previous cache model presented in [2] for LM adaptation, a new cache model is proposed in this paper. Stored in the cache are semantic components used by the dialog manager for performing progressive composition of concepts into frame structures. A continuous space LM is proposed to estimate word probabilities based on n-gram and concept histories. It is expected to perform better predictions of words expressing concepts to be composed with the already hypothesized ones even if errors in the history hypotheses may have a negative influence.

The paper introduces this new LM adaptation model and evaluates its performance on a Spoken Language Understanding (SLU) task. It is organized as follows. Section 2 summarizes previous work on LM adaptation related to the proposed approach. Section 3 introduces a new cache Neural Network LM (cacheNNLM). Section 4 reports details of experimental results of cacheNNLMs.

2. RELATED WORK

In order to take into account contexts longer than n-grams for representing contextual dependencies for word expectation, a cache memory model was proposed in [2]. Along this line, trigger models were introduced with triggers stored into a cache to predict triggered words [3]. Expectations based on the cache are combined with probabilities computed by combining general static n-grams.

The modifications of general static LM probabilities with features from the message to be analyzed were often referred to as LM adaptation. Important dimensions of LM adaptation are the type of context taken into account, how to obtain the adaptation data and how to use it to update LM probabilities.

An additional concern in LM design and adaptation is the sparseness of available data used for model parameter estimation. A possible solution to this problem is to perform history clusters after projecting vectors of word probabilities into a reduced space [4]. Neural Network LMs (NNLM) were also proposed to overcome the data sparseness problem. The NNLM [5, 6, 7] was introduced for exploiting the inherently generalization and discriminative power of a continuous vector space representation of word sequences. NNLMs are essentially approximators of functions that predict words based on histories. Words are coded in an internal network layer. An LM adaptation solution was proposed for NNLM by introducing an additional hidden layer in the network and using adaptation data to modify the weights of this layer [8].

3. THE CACHE MEMORY MODEL

An evolution of the cache memory model introduced in [2] is now proposed to perform LM adaptation using the probability of observing a linguistic event by storing in a cache memory features of the past history of the event. In most ASR applications, the considered events as well as their histories are words or word classes. Unlike for other models, as, for example, the one proposed in [4], the history features are not only words or sets of them, but also concepts expressed in the preceding dialog turns. The cache memory represents, in this way, fragments of knowledge about user intentions. Rather than modelling the attentional state as proposed in [1], the cache memory represents the history of user intentions that are also involved in the decision process of the dialog manager.

Let V_W be the word vocabulary (from the user turns) and let V_C the vocabulary of concept tags $\{C_1, \ldots, C_{|V_C|}\}$ describing knowledge chunks in a given application domain. Let σ_m indicate the sequence of words, called *support*, expressing a concept C_m in a sentence. As the corpus is annotated in terms of concepts and their supports, it is possible to create (concept, word) pairs by associating a concept tag to each word of its support, considering the sequence $(c, w)_{i,1}, \ldots, (c, w)_{i,k}, \ldots, (c, w)_{i,N(i)}$ of instances from $V_C \times V_W$, expressed in the i-th dialog turn. It will be useful to call such a pair of a concept tag expressed in the i-th dialog turn and its associated word, $(c, w)_{i,k}$, as the token $t_{i,k}$.

The problem investigated in this paper consists on modelling the influence of history tokens $t_{i-i',j}$ in the prediction of tokens in the i-th dialog turn. For simplicity, we shall omit the dialog turn when referring to the current one, so that t_k will be used instead of $t_{i,k}$. The model has to compute, for every t_k, the probability $P(t_k|h_k, h_1^{i-1})$, where h_k contains features of the tokens preceding t_k in the i-th turn, while h_1^{i-1} is a summary of the context made of the words and/or concepts expressed by the user and the machine in turns preceding the i-th turn. While the features in h_k are n-grams of tokens, h_1^{i-1} is a representation of the context proposed here for the first time. Semantic expectations depending on h_1^{i-1} and constraints represented by h_k are used to estimate the prediction probability $P(t_k|h_k, h_1^{i-1})$ computed using a cache memory model from the content of which input values are computed and applied at the input nodes of a neural network.

In practice, even if the annotated corpus is fairly large, there are not enough examples of all possible histories for estimating $P(t_k|h_k, h_1^{i-1})$. Previous work on language modelling has shown that using NNLMs as proposed in [5, 6, 7] for computing $P(w|h_k)$ is an excellent if not the best approach. For this reason, a LM adaptation is now proposed that integrates the cache model with the NNLM. The resulting model for computing $P(t_k|h_k, h_1^{i-1})$ for this task will be called *cacheNNLM*.

NNLMs can process different types of input features that can be encoded with continuous values. In this work, input neurons are fed by tokens of the current dialog turn as well as cache values computed from words and/or concepts expressed in previous turns. The possibility of using continuous values at the input makes it possible to code not only the presence of an item at the cache but also when it has been hypothesized relating the current turn. The value of a cache input neuron follows an exponential decay and is computed as a^b where $a < 1$ ($a = 0.95$ in this work) and b is the number of words to the last word or concept from previous turns. In this way, the maximum activation value is 1 and the cache activation remains the same for all tokens in the current turn. If several words or concepts activate the same neuron, only the last one is considered. The cache input layer is connected to the hidden layer as shown in Figure 1. The cacheNNLM uses both the $n-1$ n-gram history of tokens together with the cache activation to estimate the probability of the next token. Each token is coded with a 1-out-of-N value which is mapped to a lower dimension representation using a projection layer. All projection layer weights are shared (see Figure 1). In this way, both the LM and the token projections are trained together.

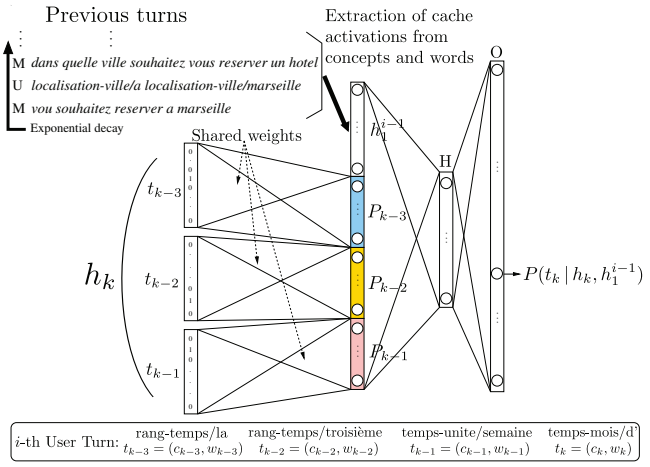

Fig. 1. Scheme of a cacheNNLM. In this example, the cache memory h_1^{i-1} of the cacheNNLM-D system, using 0.95 as exponential decay, would be: [marseille: 1.00, a: 0.95, reserver: 0.95^2, souhaitez: 0.95^3, vou: 0.95^4, localisation-ville: 0.95^5, hotel: 0.95^7, un: 0.95^8, ville: $0.95^{12}, \ldots$]. Each pair represents an input neuron in the cache memory and its value.

4. EXPERIMENTAL FRAMEWORK AND RESULTS

The proposed architecture was evaluated using the French Media corpus [9]. It was recorded using a Wizard of Oz tourist information system simulating a phone server. 1,250 dialogs were recorded, from 250 different speakers. They are manually transcribed and annotated at the concept level and are available through ELDA. The corpus is split into three

Table 1. WER on the Media corpus with different LMs.

	2-grams		*3-grams*		*4-grams*	
LM	*Dev.*	*Test*	*Dev.*	*Test*	*Dev.*	*Test*
Word n-gram	35.2	34.7	31.8	31.4	31.5	31.3
Word NNLM	31.6	31.1	30.7	30.6	30.4	30.1

subsets: a training set, a development set, and a test set containing, respectively, 12,811, 1,241, 3,468 sentences, 87,297, 9,996, 24,598 words, and 42,251, 4,652, 11,790 concept instances. The word vocabulary of the user turns V_W contains 2,007 words and the concept vocabulary V_C contains 72 different concepts. The set of observed tokens, (c, w) pairs, occurring in the training set is a small subset of all possible pairs from $V_C \times V_W$ and it is composed of $V_t = 4{,}207$ tokens. Tokens appearing only once in the training set were removed, leading to a vocabulary size $V_t' = 2{,}559$. These tokens were used for training the connectionist n-gram model.

The acoustic data have been parametrized using a window size of 30ms and 12 mel-scaled filter bands [10]. Acoustic feature vectors have been mean subtracted and divided by the standard deviation. The ASR system uses the same lexicon and phonetic transcriptions of the Speeral system [11] and left-to-right 3-state without skips context independent hybrid HMM/ANN acoustic models [12] trained with an EM algorithm based on Viterbi alignment. HMM state emissions are computed by a MultiLayer Perceptron (MLP) receiving a frame together with a left and a right context of 8 frames. The MLP contains two hidden layers of 400 neurons each one using the logistic activation function, and 105 outputs (3 states per phone) using softmax, and trained with stochastic backpropagation (sBP) using replacement. MLP training was performed using the Media training set of about 11 hours of user turns recorded at 8Khz. This corpus has been augmented with 39 hours of radio recordings of telephone conversations from the ESTER-II corpus [13].

Several word-based LMs, trained with the Media training set composed by the vocabulary of 2,007 words, were tested for the ASR task. Count-based n-gram models, with n =2, 3, and 4, were trained with SRI toolkit [14] using modified Kneser-Ney discount method. NNLMs for 2, 3, and 4-grams were also trained. The 4-gram NNLM system, which is a combination of four NNLMs and the count-based 4-gram model, provided the lowest perplexity and the WER results shown in Table 1.

The cacheNNLM is a NNLM to estimate the probability of token t_k at the i-th user turn using the $(n-1)$ previous tokens t_{k-n+1}^{k-1} and additional inputs modelling h_1^{i-1}. An additional neuron is added to each input representation of a token and to the network output to take into account the probability mass of all singletons which were removed. As a consequence, this output is divided by the total number of singletons. The size of the cache input remains small since only the most recent appearance of each concept or word appearing in the history is stored into the cache memory, following an exponential decay. Different models suitable to be used in a SLU system were tested. They are now briefly described:

- For comparison purposes, two baseline models to estimate $P(t_k|h_k)$ were tested. The input of the models is made of previous tokens $h_k = t_{k-n+1}^{k-1}$. The baseline models are:
  - baseline-a: a modified Kneser-Ney discount n-gram of tokens with the whole set of tokens.
  - baseline-b: a NNLM of non-singleton tokens linearly combined with baseline-a.
- Different cacheNNLMs extend baseline-b, to estimate $P(t_k|h_k, h_1^{i-1})$. They have the same inputs and outputs as baseline-b, h_k, plus a cache history h_1^{i-1} composed of:
  - cacheNNLM-A: concepts from previous user turns.
  - cacheNNLM-B: words from previous user turns.
  - cacheNNLM-C: concepts and words from previous user turns.
  - cacheNNLM-D: the same as cacheNNLM-C plus words from previous machine turns.

The MLPs have been trained using sBP with a weight decay regularization term, and the cross entropy error function. The projection layer has a linear activation function, hidden layer use tanh, and the softmax function is used at the output layer. In order to ensure optimal performance, every NNLM and every cacheNNLM system is a combination of four neural networks with respectively 128, 160, 192, 224 neurons in P_i and 200 neurons in the hidden layer.

4.1. Experimental results on the generation of concept hypotheses and their supports

Using the LMs considered in the previous section, it is possible to obtain the maximum likelihood sequence of tokens. In order to evaluate semantic interpretation results, it is necessary to convert each token made of a (concept, word) pair into a concept and the word sequence expressing it called support. For this purpose a Conditional Random Field (CRF) model [15] was trained with the training set using the CRF++ toolkit. Results obtained with the architecture scheme shown in Figure 2 are reported in Table 2 in terms of concept attribute error rate (CER) on concept tags, It is worth noticing that the CRFs used in this system are much simpler, using just input unigrams and bigrams, than those proposed by [16].

The improvements introduced by the cacheNNLM are statistically significant because the 95% confidence interval is 1.3 for the test set. The fact that ASR decoding generates pairs of (concept,word) hypotheses suggests to investigate in the future decoding schemes that impose combined syntactic and semantic constraints. This would inspire new more effective CRF functions that could be applied to graphs of (concept,word) hypotheses.

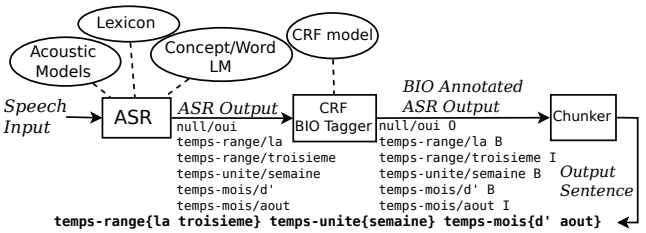

Fig. 2. Scheme of the whole system.

Table 2. Errors of concepts on the Media Development and Test sets with different LMs and word n-grams (speech input, using CRFs for concept chunking).

	2-grams		*3-grams*		*4-grams*	
Model	*Dev.*	*Test*	*Dev.*	*Test*	*Dev.*	*Test*
baseline-a	33.6	30.1	32.9	29.3	33.5	29.3
baseline-b	33.1	28.3	30.7	27.4	30.2	28.1
cacheNNLM-A	31.7	28.2	29.7	27.0	29.7	27.0
cacheNNLM-B	30.5	27.3	29.7	27.0	30.0	**26.1**
cacheNNLM-C	32.2	28.3	30.5	27.0	30.8	27.4
cacheNNLM-D	31.2	28.2	29.9	26.2	30.3	27.1

5. CONCLUSIONS AND FUTURE WORK

The use of a cache memory model is proposed in a connectionist LM which captures long term dependencies of both words and concepts. Experiments conducted on a SLU task for a spoken dialog system have shown a significant CER reduction using a rather simple model. An effective SLU module should impose semantic coherence in searching a graph of (concept,word) hypotheses, while the results reported here refer only to the 1-best hypothesis obtained with a fairly simple acoustic model. An oracle error rate for the 1-best was also computed as the percentage of concepts appearing in the manual annotation but not in the sequence obtained with different LMs. Oracle values of 15.9% for the best n-gram baseline and 14.2% for the best cacheNNLM suggests that there is plenty of room for improvement. Future work will attempt to improve the HMM/ANN acoustic models and to conceive a new SLU component acting on word-concept pairs in confusion networks.

6. ACKNOWLEDGMENTS

This work has been partially supported by the Spanish Ministry of Science and Innovation MICINN under project TIN2010-18958 (HITITA project).

7. REFERENCES

[1] M.A. Walker, "Limited attention and discourse structure," *Comput. Linguist.*, vol. 22, pp. 255–264, 1996.

[2] R. Kuhn and R. de Mori, "A cache-based natural language model for speech reproduction," *IEEE TPAMI*, vol. 12, no. 6, pp. 570–583, 1990.

[3] R. Rosenfeld, "A maximum entropy approach to adaptive statistical language modeling," *Computer Speech and Language*, vol. 10, pp. 187228, 1996.

[4] J.T. Chien and C.H. Chueh, "Dirichlet Class Language Models for Speech Recognition," *IEEE TASLP*, vol. 19, no. 3, pp. 482–495, 2011.

[5] Y. Bengio et al., "A Neural Probabilistic Language Model," *JMLR*, vol. 3, no. 2, pp. 1137–1155, 2003.

[6] M.J. Castro and F. Prat, "New Directions in Connectionist Language Modeling," in *IWANN*, 2003, pp. 598–605.

[7] Holger Schwenk, "Continuous space language models," *Comput. Speech Lang.*, vol. 21, no. 3, pp. 492–518, 2007.

[8] J. Park et al., "Improved Neural Network Based Language Modelling and Adaptation," in *Proc. Interspeech*, 2010, pp. 1041–1044.

[9] H. Bonneau-Maynard et al., "Semantic Annotation of the French Media Dialog Corpus," in *Proc. Interspeech*, 2005, pp. 3457–3460.

[10] C. Nadeu et al., "Frequency and time filtering of filter-bank energies for HMM speech recognition," in *Proc. ICSLP*, 1996, vol. 1, pp. 430–433.

[11] G. Linarès et al., "The LIA speech recognition system: from 10xRT to 1xRT," in *Proc. of TSD*. 2007, pp. 302–308, Springer-Verlag.

[12] H. Bourlard and N. Morgan, *Connectionist speech recognition—A hybrid approach*, vol. 247 of *Series in eng. and computer science*, Kluwer Academic, 1994.

[13] S. Galliano et al., "The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast New," in *Proc. Interspeech*, 2005.

[14] A. Stolcke, "SRILM: an extensible language modeling toolkit," in *Proc. ICSLP*, 2002, pp. 901–904.

[15] J. Lafferty et al., "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Machine Learning Conf.*, 2001, pp. 282–289.

[16] S. Hahn et al., "Comparing Stochastic Approaches to SLU in Multiple Languages," *IEEE TASLP*, vol. 19, no. 6, pp. 1569–1583, 2011.