Integrating Semantic information into Neural Network Language Models

Firstname Lastname

Insitute for Anthropomatics
Karlsruhe Institute of Technology, Germany
firstname.lastname@iwslt.org

Abstract

Neural models have recently shown big improvements in the performance of low-resource language modeling in phrase-based machine translation. Recurrent language models with different word factors, in particular, were a great success due to their ability to incorporate additional knowledge into the model. In this work, we want to integrate global semantic information extracted from large independent knowledge bases into neural network language models. We propose two approaches for doing this: word class extraction from Wikipedia and sentence level topic modeling. The new resulting models exhibit great potential in counteracting data scarcity problems with additional independent knowledge. This approach of integrating global context information is not restricted to language modeling but can also be easily applied to any model that profits from context or further data resources, e.g. neural machine translation. Using this model has improved rescoring quality of a state-of-the-art phrase-based translation system by ... BLEU points. We performed experiments on two language pairs.

1. Introduction

Recurrent neural network language models have recently shown great improvement in statistical machine translation, both during decoding and rescoring. The use of continuous word representations has achieved better generalizations of the data which effectively lowered data sparseness problems. Furthermore, the recurrent connections are able to model long range dependencies. Yet, most of these models strictly depend on monolingual and parallel data, which is sometimes not available in huge amounts, especially for low-resource languages. This has motivated neural network language models that take multiple parallel streams of data as input instead of just the single form of surface words. These so called factors can be used to add additional information, e.g. POS or automatic word clusters, which helps mainly with morphologically rich languages (e.g. Romanian, German). However, so far the use of factors or additional information has been limited in neural network models. Also, in those cases the extra feature only pertained to syntactic or local context knowledge around the current word. Especially for languages with low resources, it is essential to also facilitate the use of other knowledge factors, e.g. encyclopedia knowledge. It is a useful source especially for learning general concepts, even more after the emergence of the Internet has led to an explosion of textual data. These data sources give insights into a variety of human endeavors waiting to be computationally analyzed. In this paper, we study the integration of large independent knowledge bases in the form of encyclopedia, e.g. Wikipedia, into RNN-based language models and propose two solutions. First, we use the factored model to integrate extracted Wikipedia categories as one of the factors. In order to understand large unstructured datasets great achievements have been attained in latent concept learning in the area of text mining. Techniques include categorization of documents using latent semantic analysis and probabilistic topic modeling. Therefore, we employed also these techniques to compute a real-valued topic vector for each sentence that is fed into the network as additional input. Using word classes and semantic features help both sparsely inflected languages(e.g. English, Chinese) [1] as well as low-resource languages. In the model we use LSTMs to take into account both independent side information and local context information for the model prediction.

2. Related Work

Language models are a critical component of many application systems, e.g. ASR, MT and OCR. However, language models have always faced the problem of data sparseness. Factored Language Models [1] introduced the use of a bundle of factors associated with a word which outperformed previous n-gram models without expanding the training data. For factors morphs, stems, POS and word class obtained using the SRILM's n-gram-class tool were used. [2] replaced the single feature stream of surface words with multiple factors and integrated it into phrase-based statistical machine translation systems by breaking down the translation model into several steps that pertain to the translation of single factors which are all taken into account when the target word is generated. After recurrent neural network models became a success in language modeling [3], a factored input layer was employed in a model by [4] which uses a structured output layer based on word classes that was able to handle vocabulary of arbitrary size. Motivated by multi-task learning in NLP, [5] proposed a multi-factor recurrent neural network language model which jointly predicts different output factors by mapping the output of the LSTM-layer to as many softmax layers as there are output factors, thus creating multiple distributions at the output layer. In the rescoring of an n-best list, this model can be included as either one additional feature or several features depending on whether the output is treated as a joint probability or individual probabilities. One disadvantage of factored input is that additional factors must match the surface words in space. As a consequence, surface words that uses 1-of-n encoding cannot have factors with continuous space representations. However, this is often necessary to model more complex structures, e.g. topic distributions. In [6], a topic-conditioned RNNLM is proposed which takes a real-valued input vector as an additional input in association with each word. This vector is used to convey local context information based on previous sentences using latent dirichlet allocation. Often, the meaning of a word cannot be just derived from its preceding words but by content words in the entire sentence or surrounding sentences. However, the model cannot take side information associated with a sentence that contains the current word, which is what we studied in the second part.

3. Integration of Side Information

3.1. Word class extraction

3.2. Sentence level topic modeling

4. Experiments

4.1. System Description

4.2. English-Chinese

4.3. English-Romanian

5. Conclusion

6. Acknowledgements

7. References

[1] J. A. Bilmes and K. Kirchhoff, "Factored language models and generalized parallel backoff," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers-Volume 2*. Association for Computational Linguistics, 2003, pp. 4–6.

[2] P. Koehn and H. Hoang, "Factored translation models." in *EMNLP-CoNLL*, 2007, pp. 868–876.

[3] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, "Recurrent neural network based language model." in *Interspeech*, vol. 2, 2010, p. 3.

[4] Y. Wu, X. Lu, H. Yamamoto, S. Matsuda, C. Hori, and H. Kashioka, "Factored language model based on recurrent neural network," 2012.

[5] J. Niehues, T.-L. Ha, E. Cho, and A. Waibel, "Using factored word representation in neural network language models."

[6] T. Mikolov and G. Zweig, "Context dependent recurrent neural network language model." in *SLT*, 2012, pp. 234–239.