Integrating Semantic information into Neural Network Language Models

Firstname Lastname

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

Abstract

Neural network and translation models have recently shown great potentials in improving the performance of 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. At the same time, great achievements have been attained in latent concept learning in the area of text mining. In this work, we combined both ideas to integrate global semantic information extracted from large independent knowledge bases into neural network language models. We propose two novel approaches for doing this: word class extraction from Wikipedia and sentence level topic modeling. The new resulting models are especially helpful for morphologically rich languages without the need to expand the training corpus but also in solving lexical ambiguities. 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, e.g. neural network 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 reason for them outperforming previous language models is that they maintain a hidden layer with recurrent connections to their previous values which makes them able to capture long span dependencies. However, often we can not control which relations are learned by the network, which motivated the attempt to provide neural networks with useful input information in the first place. 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, thus generalizing the input and output layer of conventional models. 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) but can also consider even longer context and therefore improve the modeling of the overall structure. 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 attributed to syntactic or local context knowledge around the current word. 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. This is motivated by the emergence of the Internet that has led to an explosion of textual data in particular. These data sources give insights into a variety of human endeavors waiting to be computationally analyzed. First, we propose the novel idea to use extracted Wikipedia categories as one of the word factors in factored neural network models. 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. As a second approach, we employed these in the RNNLM in this work by explicitly computing a background topic vector for each sentence that is fed into the network as additional input. Using word classes and semantic features helps with sparsely inflected languages(e.g. English, Chinese) [1] In the model we use LSTMs to consider 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], factored input layer was employed by [4] which used 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 one additional feature or several features depending on if 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 word data 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 function 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

The paper title must be in boldface. All non-function words must be capitalized, and all other words in the title must be lower case. The paper title is centered across the top of the two columns on the first page as indicated above.

3.2. Sentence level topic modeling

The authors' name(s) and affiliation(s) appear centered below the paper title. If space permits, include a mailing address here. The templates indicate the area where the title and author information should go. These items need not be confined to the number of lines indicated; papers with multiple authors and affiliations may require two or more lines. Note that the submission version of technical papers *should be anonymized for review*.

4. Experiments

4.1. System Description

4.2. English-Chinese

4.3. English-Romanian

5. Conclusion

6. Acknowledgements

The IWSLT 2015 organizing committee would like to thank the organizing committees of INTERSPEECH 2004 for their help and for kindly providing the template files.

- Proceedings will be printed in A4 format. The layout is designed so that files, when printed in US Letter format, include all material but margins are not symmetric. Although this is not an absolute requirement, if at all possible, **PLEASE TRY TO MAKE YOUR SUBMISSION IN A4 FORMAT.**
- Two columns are used except for the title part and possibly for large figures that need a full page width.
- Left margin is 20 mm.
- Column width is 80 mm.
- Spacing between columns is 10 mm.
- Top margin 25 mm (except first page 30 mm to title top).
- Text height (without headers and footers) is maximum 235 mm.
- Headers and footers must be left empty (they will be added for printing).
- Check indentations and spacings by comparing to this example file (in pdf format).

6.0.1. Headings

Section headings are centered in boldface with the first word capitalized and the rest of the heading in lower case. Subheadings appear like major headings, except they start at the left margin in the column. Sub-sub-headings appear like subheadings, except they are in italics and not boldface. See the examples given in this file. No more than 3 levels of headings should be used.

6.1. Text font

Times or Times Roman font is used for the main text. Recommended font size is 9 points which is also the minimum allowed size. Other font types may be used if needed for special purposes. While making the final PostScript file, remember to include all fonts!

LaTeX users: DO NOT USE Computer Modern FONT FOR TEXT (Times is specified in the style file). If possible, make the final document using POSTSCRIPT FONTS. This is necessary given that, for example, equations with non-ps Computer Modern are very hard to read on screen.

6.2. Figures

All figures must be centered on the column (or page, if the figure spans both columns). Figure captions should follow each figure and have the format given in Fig. 1.

Figures should preferably be line drawings. If they contain gray levels or colors, they should be checked to print well on a high-quality non-color laser printer.

Table 1: *This is an example of a table.*

ratio	decibels
1/1	0
2/1	≈ 6
3.16	10
10/1	20
1/10	-20

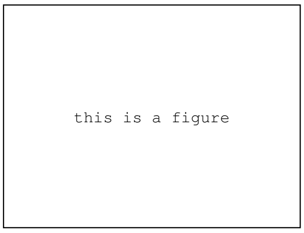

Figure 1: *Schematic diagram of speech production.*

6.3. Tables

An example of a table is shown as Table 1. Somewhat different styles are allowed according to the type and purpose of the table. The caption text may be above or below the table.

6.4. Equations

Equations should be placed on separate lines and numbered. Examples of equations are given below. Particularly,

$$x(t) = s(f_\omega(t)) \tag{1}$$

where $f_\omega(t)$ is a special warping function

$$f_\omega(t) = \frac{1}{2\pi j} \oint_C \frac{\nu^{-1k} d\nu}{(1 - \beta\nu^{-1})(\nu^{-1} - \beta)} \tag{2}$$

A residue theorem states that

$$\oint_C F(z)dz = 2\pi j \sum_k Res[F(z), p_k] \tag{3}$$

Applying (3) to (1), it is straightforward to see that

$$1 + 1 = \pi \tag{4}$$

Make sure to use `\eqref` when refering to equation numbers. Finally we have proven the secret theorem of all speech sciences (see equation (3) above). No more math is needed to show how useful the result is!

6.5. Hyperlinks

Hyperlinks can be included in your paper. Moreover, be aware that the paper submission procedure includes the option of specifying a hyperlink for additional information. This hyperlink will be included in the CD-ROM. Particularly pay attention to the possibility, from this single hyperlink, to have further links to information such as other related documents, sound or multimedia.

If you choose to use active hyperlinks in your paper, please make sure that they present no problems in printing to paper.

6.6. Page numbering

Final page numbers will be added later to the document electronically. *Please don't make any headers or footers!*.

6.7. References

The reference format is the standard for IEEE publications. References should be numbered in order of appearance, for example [**?**], [**?**], and [**?**].

7. Experiments

Please make sure to give all the necessary details regarding your experimental setting so as to ensure that your results could be reproduced by other teams.

8. Conclusions

This paper has described a novel approach for doing wonderful stuff such as ...

9. 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.