Learning Word Representations

2nd Practical, Unsupervised Language Learning

18 April 2018

1 Introduction

In the first practical, you evaluated three different pre-trained word embeddings. Now, you will implement and train three different models to learn the word embeddings. You will implement 3 models of word representation, one trained for maximum likelihood, and two latent variable models trained by variational inference. The word representation learning models that you will implement are: The skip-gram [4], the Bayesian skip-gram [1], and Embed-Align [5]. Skip-gram is trained discriminatively by having a central word predict context words in a window surrounding it. Bayesian skip-gram introduces stochastic latent embeddings, but does not change the discriminative nature of the training procedure. Embed-Align introduces stochastic latent embeddings as well as a latent alignment variable and learns by generating translation data. Eventually, you should compare the performance of these three models on the lexical substitution task.

2 Skip-gram

In the Skip-gram architecture, we use a central/target word to predict context words. This is in contrast with the Continuous Bag of Word (CBoW) where we use the context to predict the target word. Both skip-gram and CBoW are introduced in [4].

3 Bayesian Skip-gram

In the normal skip-gram model, the words are represented as deterministic vectors in a low dimensional space. One of the shortcomings of this type of word embeddings is that they are independent from the dynamic context in which a word occurs. In situations where a word can have different meanings in different context, they are not able to distinguish between the different senses of the words. To address this issue, in the Bayesian Skip-gram model (BSG) the choice of context words is dependent on the context-specific latent representation of the central word [1]. In this model we learn to represent words as Gaussian probability densities, where the densities reflect the distributions over the possible representations.

4 Embed-Align

The Embed-Align model is a generative model of word representation that learns from positive correlations implicitly expressed in translation data. In this model, we learn word representations by learning their lexical translations in a foreign language. It does not require a notion of central window, instead, it requires sentence-aligned parallel data. Similar to BSG, words are represented as Gaussian densities and are sensitive to context.

5 Datasets to train the models

Skip-gram and BSG are trained on monolingual data, while EmbedAlign requires bilingual data. Below we list the resources available.

For training we provide two parallel corpora (a small one for development and debugging, and a larger one which you should use for experiments and report). For monolingual models you should use only the English portion of the dataset, for Embed-Align you should consider English as L_1 and French as L_2 (see terminology of [5]).

- Small collection: already pre-processed English-French hansards.
  https://surfdrive.surf.nl/files/index.php/s/SZmadxD7QTwEPx0
- Large collection: already pre-processed English-French Europarl
  https://surfdrive.surf.nl/files/index.php/s/Bliv4tIwd7NLAxP
- Gold-standard alignments: used for dev and test (relevant for validation and test of Embed-Align only).
  https://surfdrive.surf.nl/files/index.php/s/C4QRRulMMX4bdhn

6 Evaluation: Lexical Substitution Task

The Lexical Substitution Task (LST) is introduced in [3]. The task is: Finding alternative words/phrases that can replace a target word. In other words, for a given sentence, and a target word in the sentence, we need to rank all other words as candidates for replacing the target word.

6.1 Dataset

The LST dataset includes 201 target words present in 10 sentences/contexts each, along with a manually annotated list of potential replacements. The data are split in 300 instances for validation and 1,710 for test. You can download the dataset for LST, along with the evaluation scripts, from here:

- https://surfdrive.surf.nl/files/index.php/s/71bLDwNbeTOX1IA

6.2 Evaluation Metric

You should report the performances of the models in terms of Generalized Averaged Precision (GAP). GAP compares a set of predicted rankings to a set of gold standard rankings [2]. For Embed-Align you should also report performance in terms of alignment error rate (AER).

7 Write a report

Write a 1 page report, briefly describing

- the methods you have implemented
- the experiments you have conducted
- the results of these experiments
- what you have learned about the properties of the three word representation models

Submit your report in a pdf format on blackboard with the title **ULL-Practical1-FullName1_FullName2**.

Submission deadline for the report is **23:59 on Thursday, 10 May**. Submit only one report per group. Upload your codes on a github repository and put a link to the repository in your report. Also, add the instructions of how to run your codes for each part of the practical to the github ReadMe.

References

[1] Arthur Bražinskas, Serhii Havrylov, and Ivan Titov. Embedding words as distributions with a bayesian skip-gram model. *arXiv preprint arXiv:1711.11027*, 2017.

[2] Kazuaki Kishida. *Property of average precision and its generalization: An examination of evaluation indicator for information retrieval experiments.* National Institute of Informatics Tokyo, Japan, 2005.

[3] Diana McCarthy and Roberto Navigli. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 48–53. Association for Computational Linguistics, 2007.

[4] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[5] Miguel Rios, Wilker Aziz, and Khalil Sima'an. Deep generative model for joint alignment and word representation. *arXiv preprint arXiv:1802.05883*, 2018.