ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Broad Discourse Context for Language Modeling

Master Thesis

Moisés Torres

November 1, 2017

Supervisor:
Prof. Dr. Thomas Hofmann

Co-supervisors:
Florian Schmidt
Paulina Grnarova

Department of Computer Science, ETH Zürich

Abstract

First paragraph, high level description of the problem. Check project proposal!

Describe thesis approach, mention lambada maybe

Acknowledgments

I would like to thank... Finish acknowledgements

Contents

Chapter 1

Introduction

Here comes the intro...

Finish intro (3 pages)

1.1 Problem Statement and Motivation

motivation...

Finish motivation (1 page)

1.2 Thesis Contributions

contributions...

Finish contributions (1 page)

1.3 Thesis Outline

outline...

Finish outline (1 page)

Chapter 2

Related Work

This chapter will describe [1] the state-of-the-art [2] of...

Finish related work (3 pages)

Chapter 3

Neural Language Modeling

In this chapter, we introduce the notation used throughout the thesis and give a brief overview of the development of neural language modeling (NLM) since its inception. We also review some of the main weaknesses shown by this family of models and how they have been addressed in the literature.

- Finish this (7-8 pages)
- Mention CNN LMs ?
- Softmax variants (hierarchical)?
- Modern models (highway)?
- Include references for n-grams?
- Figures?

3.1 Notation

Before continuing, we will define the notation used in this thesis:

- Scalars are denoted with lowercase letters, such as x.
- Vectors (of size N) are denoted with bold lowercase letters, such as $\mathbf{x}$ with its i-th element $\mathbf{x}_i$, and are always assumed to be column vectors.
- Matrices (of size $N \times M$) are denoted with uppercase letters, such as X with X_{ij} as its (i, j)-th element.

3.2 Background

Prior to introducing the specifics of NLMs, we will formalize the task at hand and introduce some of its core concepts.

3.2.1 Language Modeling

First, we define a **word-based language model** as a model able to compute the probability of a sentence or sequence words $P(w_1, \ldots, w_n)$. Such models are of great use in tasks where we have to recognize words in noisy or ambiguous input such as speech recognition or machine translation, among others.

If now we decompose the joint probability of a sequence using the chain rule of probability as shown in Equation 3.1, we observe that the function that needs to be estimated boils down to the conditional probability of a word given the history of previous words. However, taking into account the whole context poses a problem as language is creative and any particular sequence might have occurred few (or no) times before. Many of the models that we will introduce opt to approximate the real conditional distribution by making a Markov assumption as shown in Equation 3.2. This means that the probability of an upcoming word is fully characterized by the $n-1$ previous ones. Despite seeming an incorrect assumption for a complex source of information such as language, it has been proven to work really well in practice.

$$\begin{aligned} P(w_1, \ldots, w_n) &= P(w_1)P(w_2|w_1)P(w_3|w_1^2)\ldots P(w_n|w_1^{n-1}) \\ &= \prod_{k=1}^{n} P(w_k|w_1^{k-1}) \end{aligned} \tag{3.1}$$

$$P(w_k|w_1^{k-1}) \approx P(w_k|w_{k-1}^{k-n}) \tag{3.2}$$

3.2.2 Evaluation

Following a common practice in machine learning, we use a test set in order to evaluate our models. In the case of language modeling we have a word sequence $W_1^n = \{w_1, \ldots, w_n\}$ and the better the model is, the higher the probability it will assign to this sequence. Rather than working directly with raw probabilities we define a metric called **perplexity**, which is the geometric average of the inverse of the probability over the test set, as shown in Equation 3.3. Therefore, lower perplexity is better.

$$\begin{aligned} \text{Perplexity}(W_1^n) = P(W_1^n)^{-\frac{1}{n}} &= \sqrt[n]{\frac{1}{P(W_1^n)}} \\ &= \sqrt[n]{\frac{1}{\prod_{k=1}^{n} P(w_k|W_1^{k-1})}} \end{aligned} \tag{3.3}$$

Moreover, we can regard language as a source of information and apply the Information Theory toolbox to find a different (and equivalent) interpretation of perplexity. For that we need to introduce the basic concept of **entropy** (Equation 3.4 shows its formulation for discrete variables), which measures the expected uncertainty or "surprise" S of the value of a random variable X. Without going into details, it is easy to see that defining uncertainty as the negative logarithm (the specific base doesn't matter, but traditionally it

is assumed to be 2) of the probability of each event matches our intuition (like $S(p) > S(q)$ then $p < q$).

$$H(X) = \mathbb{E}[S(X)] = -\sum_{x \in \mathcal{X}} P(x) \log_2(P(x)) \quad \text{with} \quad S(\cdot) = -\log_2(\cdot) \tag{3.4}$$

A difference when it comes to language is that it involves dealing with sequences W_1^n of discrete random variables. For a given language L we can define the entropy of a variable ranging over all possible sequences of length n. To obtain the entropy-per-word we would only need to normalize by n (Equation 3.5).

$$\frac{1}{n} H(W_1^n) = -\frac{1}{n} \sum_{W_1^n \in L} P(W_1^n) \log_2(P(W_1^n)) \tag{3.5}$$

However, in order to calculate the true entropy of a language we need to consider sequences of infinite length (Equation 3.6). Shannon-McMillan-Breiman theorem states that if the language is regular in certain ways we can take a single long enough sequence instead of summing over all possible sequences (* in Equation 3.6).

$$\begin{aligned} H(L) &= -\lim_{n \to \infty} \frac{1}{n} \sum_{W_1^n \in L} P(W_1^n) \log_2(P(W_1^n)) \\ &\overset{*}{=} -\lim_{n \to \infty} \frac{1}{n} \log_2(P(W_1^n)) \end{aligned} \tag{3.6}$$

Similarly we have **cross-entropy** which measures the relative entropy of P with respect to M, P being the true probability distribution and M a model (e.g. an approximation) of P. After applying Shannon-McMillan-Breiman theorem and assuming that n is large enough, we can see in Equation 3.7 the final formulation of the cross-entropy, which is used as the default loss function when optimizing neural language models.

$$\begin{aligned} H(P, M) &= -\lim_{n \to \infty} \frac{1}{n} \sum_{W_1^n \in L} P(W_1^n) \log_2(M(W_1^n)) \\ &\overset{*}{=} -\lim_{n \to \infty} \frac{1}{n} \log_2(M(W_1^n)) \approx -\frac{1}{n} \log_2(M(W_1^n)) \\ &= -\frac{1}{n} \sum_{k=1}^{n} \log_2(M(w_k | W_1^{k-1})) \end{aligned} \tag{3.7}$$

Finally, we can see in Equation 3.8 how cross-entropy and perplexity are connected. This relation gives raise to a nice interpretation of perplexity as

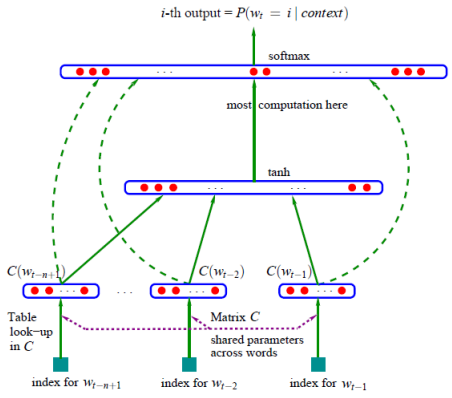
branching factor: entropy measures uncertainty (in bits, if we use $\log_2$) but in exponentiated form it's measured as the cardinality of a uniform distribution with equivalent uncertainty.

$$\text{Perplexity}(W_1^n) = 2^{\text{cross-entropy}} = 2^{H(P,M)} = M(W_1^n)^{-\frac{1}{n}} \tag{3.8}$$

3.3 Feed-Forward Neural Language Models (FFNLM)

Until the appearance of NLMs the most successful approaches were based on n-grams, which are Markov models that estimate words from a fixed window of previous words and estimate probabilities by counting in a corpus and normalizing. Due to their nature n-gram estimates intrinsically suffer from sparsity and several methods like smoothing, backoff and interpolation have been proposed to deal with this problem.

Along those lines, the first successful attempt of applying neural networks [3] raised the point that when modeling the joint distribution between many discrete random variables (such as words in a sentence), any change of these variables may have a drastic impact on the value of the estimated function. On the contrary, by using continuous variables we obtain better generalization because the function to be learned can be expected to have some local smoothness properties ("similar" words should get similar probabilities). While taking longer to train, this approach is able to achieve significantly better results by jointly learning word representations and a statistical language model.



$$\begin{aligned} \mathbf{x} &= [C(w_{t-1}), C(w_{t-2}), \ldots, C(w_{t-n+1})] \\ \mathbf{y} &= W\mathbf{x} + U\tanh(H\mathbf{x} + \mathbf{d}) + \mathbf{b} \\ P(w_t = i | w_{t-1}, \ldots, w_{t-n+1}) &= \frac{e^{\mathbf{y}_i}}{\sum_n e^{\mathbf{y}_n}} \end{aligned} \tag{3.9}$$

Figure 3.1: Feed-Forward NLM architecture

Similar to n-grams, the model introduced in [3] conditions the probability of a word on the previous $n-1$ words. The main difference lies in the

concept of "distributed feature vectors"; words are embedded into a vector-space by assigning them a continuous real-vector representation. As seen in Figure 3.1, this is done via a look-up operation over the embedding matrix C. The concatenated word representations are then fed through one or more hidden layers and the resulting hidden representation is used to generate the unscaled log probabilities with a fully connected layer. Finally, a softmax operation produces a valid probability distribution over the full vocabulary.

3.4 Word Vectors

As we have seen in the previous section, distributed continuous vectors allow for "clever" smoothing by taking into account automatically learnt syntactic and semantic features. [4] picked up on this concept trying to find ways of training these vector representations more efficiently. The paper introduces a family of models known as **word2vec**, whose architecture matches the one from a FFNLM where the non-linear hidden layer has been removed (and we end up with a simple log bilinear model). The difference between them lies on the "fake" objective (fake in the sense that we are only interested in the resulting embeddings and not the actual outputs of the model) they optimize for learning the word embeddings:

- Continuous Bag-of-Words model (CBOW): given a symmetric window of size k around a specific position $\{w_{i-k}, \dots, w_{i-1}, w_{i+1}, \dots w_{i+k}\}$ we want to predict that word w_i. The term "bag-of-words" comes from the fact that the embeddings of the whole window are summed (instead of concatenated) and thus, order is not kept anymore.
- Continuous Skip-Gram model: given a specific position, we randomly sample words inside its surrounding window and try to predict them. Therefore, each training example is a tuple consisting of w_i as input and a word from the window as output.

In addition to a simplified architecture and a modified objective, further optimizations for the Skip-Gram model were introduced in the follow-up paper [5]. As we already saw in Equation 3.9??, most of the computation is done in the softmax operation over the full vocabulary. In order to avoid this, we cast our task to a binary classification problem by making use of a new objective called **negative sampling**. Inspired by noise contrastive estimation (NCE), the task is to distinguish the target word w_O from draws from a noise distribution $P_n(w)$ (e.g. unigram distribution) using logistic regression, where there are k negative samples for each data sample (Equation 3.10).

$$\mathcal{L}(\theta) = \log(P(w_O|w_I)) = \log(\sigma(\mathbf{v}_{w_O}^\top \mathbf{v}_{w_I})) + \sum_{i=1}^{k} \log(-\sigma(\mathbf{v}_{w_i}^\top \mathbf{v}_{w_I})) \quad \text{with} \quad w_i \sim P_n(w) \tag{3.10}$$

Another famous family of word vectors is **GloVe** [6], where the objective is a weighted (weighting function $f(\cdot)$) least squares fit of the log-counts (Equation 3.11). Rather than taking a predictive model approach (like word2vec) to learn their vectors in order to improve their predictive ability, GloVe does dimensionality reduction on the co-occurrence counts matrix N.

$$\mathcal{L}(\theta, N) = \sum_{i,j:N_{ij}>0} f(N_{ij})(\log(N_{ij}) - (\mathbf{v}_{w_O}^\top \mathbf{v}_{w_I} + b_O + b_I))^2 \tag{3.11}$$

In summary, word vectors have become a standard in NLP and are used as input in all sorts of downstream tasks such as sentiment analysis.

3.5 Recurrent Neural Language Models (RNNLM)

he [7] as in

$$\mathbf{h}_t = \sigma(W\mathbf{x}_t + U\mathbf{h}_{t-1}) \tag{3.12}$$

$$\begin{aligned} \mathbf{f}_t &= \sigma_g(W_f\mathbf{h}_{t-1} + U_f\mathbf{x}_t + \mathbf{b}_f) \\ \mathbf{i}_t &= \sigma_g(W_i\mathbf{h}_{t-1} + U_i\mathbf{x}_t + \mathbf{b}_i) \\ \mathbf{o}_t &= \sigma_g(W_o\mathbf{h}_{t-1} + U_o\mathbf{x}_t + \mathbf{b}_o) \\ \tilde{\mathbf{c}}_t &= \sigma_c(W_c\mathbf{h}_{t-1} + U_c\mathbf{x}_t + \mathbf{b}_c) \\ \mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t \\ \mathbf{h}_t &= \mathbf{o}_t \odot \sigma_h(\mathbf{c}_t) \end{aligned} \tag{3.13}$$

Chapter 4

Rare Word Prediction

bla bla

Finish this

$$\mathcal{L} = -\sum_{j} \hat{y_{ij}} \log(p(y_{ij}|x_i)) - \log(g + \sum_{i \in I(y,x)} a_i) \tag{4.1}$$

Chapter 5

Experiments and Results

experiments...

Finish experiments and results

$$\mathcal{L} = -\sum_{n} \log(P(name|h_t)P(w_t|h_t, name) + (1 - P(name|h_t))P(w_t|h_t, notName)) \\ +\lambda(y_{name} \log(P(name|h_t)) + (1 - y_{name}) \log(1 - P(name|h_t)) \tag{5.1}$$

		All		Names	
		Train	Dev	Train	Dev
Basic	Baseline	42.5	61.5	1000	130K
	Mixture ($\lambda = 100$)	55	69	4000	120K
Input dropout	Baseline	52.5	63	2000	140K
	Mixture ($\lambda = 100$)	65	72	8000	120K
State dropout	Baseline	48	62.5	1000	120K
	Mixture ($\lambda = 100$)	62.5	73	6000	120K
Output dropout	Baseline	62.5	65	5000	150K
	Mixture ($\lambda = 100$)	80	73	20K	80K
L2 regularization ($\beta = 0.01$)	Baseline	100	125	10K	400K
	Mixture ($\lambda = 100$)	107.5	119	8000	120K

Table 5.1: Perplexity

Chapter 6

Conclusion

brief remarks...

Finish conclusion (1/2 page)

6.1 Achieved Results

results...

Finish achieve results (1/2 page)

6.2 Future Work

future work...

Finish future work (1/2 page)

Bibliography

[1] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. Regularizing and optimizing lstm language models. *arXiv preprint arXiv:1708.02182*, 2017.

[2] Edouard Grave, Armand Joulin, and Nicolas Usunier. Improving neural language models with a continuous cache. *arXiv preprint arXiv:1612.04426*, 2016.

[3] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

[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] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[6] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[7] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[8] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Interspeech*, volume 2, page 3, 2010.

[9] Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The lambada dataset: Word prediction requiring a broad discourse context. *arXiv preprint arXiv:1606.06031*, 2016.

[10] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*, 2017.

[11] Michał Daniluk, Tim Rocktäschel, Johannes Welbl, and Sebastian Riedel. Frustratingly short attention spans in neural language modeling. *arXiv preprint arXiv:1702.04521*, 2017.

[12] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

[13] Caglar Gulcehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio. Pointing the unknown words. *arXiv preprint arXiv:1603.08148*, 2016.

[14] Zichao Yang, Phil Blunsom, Chris Dyer, and Wang Ling. Reference-aware language models. *arXiv preprint arXiv:1611.01628*, 2016.

[15] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014.

[16] Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in neural information processing systems*, pages 1019–1027, 2016.

[17] Stanislau Semeniuta, Aliaksei Severyn, and Erhardt Barth. Recurrent dropout without memory loss. *arXiv preprint arXiv:1603.05118*, 2016.

[18] Hakan Inan, Khashayar Khosravi, and Richard Socher. Tying word vectors and word classifiers: A loss framework for language modeling. *arXiv preprint arXiv:1611.01462*, 2016.

[19] Daniel Jurafsky and James H. Martin. *Speech and Language Processing (2Nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2009.

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

Authored by (in block letters):
For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.