Learning to Explain Non-Standard English Words and Phrases

Ke Ni and **William Yang Wang**
Department of Computer Science
University of California, Santa Barbara
Santa Barbara, CA 93106 USA
{ke00@umail},{william@cs}.ucsb.edu

Abstract

We describe a data-driven approach for automatically explaining new, non-standard English expressions in a given sentence, building on a large dataset that includes 15 years of crowdsourced examples from *UrbanDictionary.com*. Unlike prior studies that focus on matching keywords from a slang dictionary, we investigate the possibility of learning a neural sequence-to-sequence model that generates explanations of unseen non-standard English expressions given context. We propose a dual encoder approach—a word-level encoder learns the representation of context, and a second character-level encoder to learn the hidden representation of the target non-standard expression. Our model can produce reasonable definitions of new non-standard English expressions given their context with certain confidence.

1 Introduction

In the past two decades, the majority of NLP research focused on developing tools for the Standard English on newswire data. However, the non-standard part of the language is not well-studied in the community, even though it becomes more and more important in the era of social media. While we agree that one must take a cautious approach to automatic generation of non-standard language (Hickman, 2013), but for many practical purposes, it is also of crucial importance for machines to be able to *understand* and *explain* this important subversion of the language.

In the NLP community, using dictionaries of non-standard language as an external knowledge source is useful for many tasks. For example, Burfoot and Baldwin (2009) consult the slang definitions from Wiktionary to detect satirical news articles. Wang and McKeown (2010) show that using a slang dictionary of 5K terms can help detecting vandalism of Wikipedia edits. Han and Baldwin (2011) make use of the same slang dictionary, and achieve the best performance by combining the dictionary lookup approach with word similarity and context for Twitter and SMS text normalization. However, using a 5K slang dictionary may suffer from the coverage issue, since slang is evolving rapidly in the social media age[1]. Recently, Thanapon Noraset (2016) shows that it is possible to use word embeddings to generate plausible definitions. Nonetheless, one weakness is that definition of a word may change within different contexts.

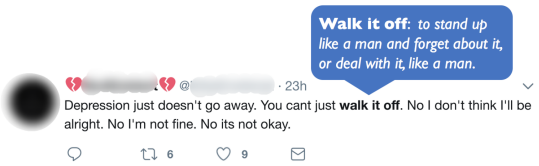

Figure 1: An example Tweet with a non-standard English expression. Our model aims at automatically explaining any newly emerged, non-standard expressions (generating the blue box).

In contrast, we take a more radical approach: we aim at building a general purpose sequence-to-sequence neural network model (Sutskever et al., 2014) to explain any non-standard English expression, which can be used in many NLP and social media applications. More specifically, given a sentence that includes a non-standard English expression, we aim at automatically generating the translation of the target expression. Previously, this is not possible because the resources of labeled

[1]For example, more than 2K entries are submitted daily to Urban Dictionary (Kolt and Lazier, 2009), the largest online slang resource.

non-standard expressions are not available. In this paper, we collect a large corpus of 15 years of crowdsourced examples, and formulate the task as a sequence-to-sequence generation problem. Using a word-level model, we show that it is possible to build a general purpose non-standard English words and phrases explainer using neural sequence learning techniques. To summarize our contributions:

- We present a large, publicly available corpus of non-standard English words and phrases, including 15 years of definitions and examples for each entry via crowdsourcing;
- We present a hybrid word-character sequence-to-sequence model that directly explains unseen non-standard expressions from social media;
- Our novel dual encoder LSTM model outperforms a standard attentive LSTM baseline, and it is capable of generative plausible explanation for unseen non-standard words and phrases.

In the next section, we outline related work on non-standard expressions and social media text processing. We will then introduce our dual encoder based attentive sequence-to-sequence model for explaining non-standard expressions in Section 3. Experimental results are shown in Section 4. And finally, we conclude in Section 5.

2 Related Work

The study of non-standard language is of interests to many researchers in the social media and NLP communities. For example, Eisenstein et al. (2010) propose a latent variable model to study the lexical variation of the language on Twitter, where many regional slang words are discovered. Zappavigna (2012) identifies the Internet slang as an important component in the discourse of Twitter and social media. Gouws et al. (2011) provide a contextual analysis of how social media users shorten their messages. Notably, a study on Tweet normalization (Han and Baldwin, 2011) finds that, even when using a small slang dictionary of 5K words, Slang makes up 12% of the ill-formed words in a Twitter corpus of 449 posts. In the NLP community, slang dictionary is widely used in many tasks and applications (Burfoot and Baldwin, 2009; Wang and McKeown, 2010; Rosenthal and McKeown, 2011). However, we argue that using a small, fixed-size dictionary approach may be suboptimal: it suffers from the low coverage problem, and to keep the dictionary up to date, maintaining such dictionary is also expensive and time-consuming. To the best of our knowledge, our work is the first to build a general purpose machine learning model for explaining non-standard English terms, using a large crowdsourced dataset.

3 Our Approach

3.1 Sequence-to-Sequence Model

Since our goal is to automatically generate explanations for any non-standard English expressions, we select sequence-to-sequence models with attention mechanisms as our fundamental framework (Bahdanau et al., 2014), which can produce abstractive explanations, and assign different weights to different parts of a sentence. To model both the context words and the non-standard expression, we propose a hybrid word-character dual encoder. An overview of our model is shown in Figure 2.

3.2 Context Encoder

Our context encoder is basically a recurrent neural network with long-short term memory (LSTM) (Hochreiter and Schmidhuber, 1997). LSTM consists of input gates, forget gates and output gates, together with hidden units as its internal memory. Here, i controls the impact of new input, while f is a forget gate, and o is an output gate. $\tilde{C}_t$ is the candidate new value. h is the hidden state, and m is the cell memory state. "$\odot$" means element-wise vector product. The definition of the gates, memory, and hidden state is:

$$i_t = \sigma(W_i[x_t, h_{t-1}])$$

$$f_t = \sigma(W_f[x_t, h_{t-1}])$$

$$o_t = \sigma(W_o[x_t, h_{t-1}])$$

$$\tilde{C}_t = \tanh(W_c[x_t, h_{t-1}])$$

$$m_t = m_{t-1} \odot f_t + i_t \odot \tilde{C}_t$$

$$h_t = m_t \odot o_t$$

At each step, RNN is given a vector as input, changes its hidden states and produces outputs from its last layer. Hidden states and outputs are stored and later passed to a decoder, which produces final outputs based on hidden states, outputs,

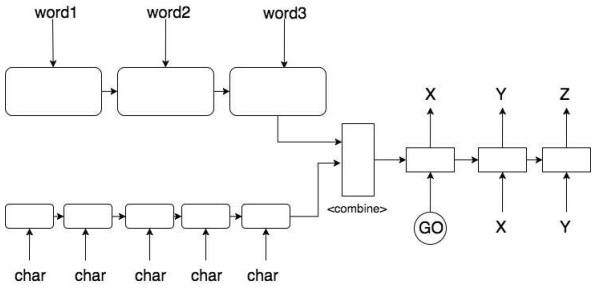

Figure 2: Dual Encoder Structure. *Top left: a word-level LSTM encoder for modeling context. Bottom left: a character-level LSTM encoder for modeling target non-standard expression. Right: an LSTM decoder for generating explanations.*

and the decoder's own parameters.The context encoder learns and encodes sentence-level information for the decoder to generate explanations.

3.3 Attention Mechanism

The idea of designing an attention mechanism is to select key focuses in the sentence. More specifically, we would like to pay attention to specific parts of encoder outputs and states. We follow a recent study (Vinyals et al., 2015) to setup an attention mechanism. We have a separate LSTM for decoding.

Here we briefly explain the method. When the decoder starts its decoding process, it has all hidden units from the encoder, denoted as $(h_1..h_{T_1})$. Also we denote the hidden state of the decoder as $(d_1..d_{T_2})$. T_1 and T_2 are input and output lengths. At each time step, the model computes new weighted hidden states based on encoder states and three learnable components: v, W_1' and W_2'.

$$u_i^t = v^T tanh(W_1' h_i + W_2' d_t)$$

$$a_i^t = softmax(u_i^t)$$

$$d_t' = \sum_{i=1}^{T_1} a_i^t h_i$$

Here a denotes the attention weights. u^t has length T_1. After the model computes d_t', it concatenates d_t' and d_t as new hidden states used for prediction and next update.

3.4 Dual Encoder Structure

Given a single context encoder, it is challenging for any decoder to generate explanation in the instance. The reason is that there could be multiple non-standard expressions in the sentence, and it confuses the decoder on which one it should explain. In practice, the user can often pin point to the exact expression she or he does not know, so in this work, we design a second encoder to learn the representation of the target non-standard expression.

Since our goal is to explain any non-standard English words and phrases, we consider a character-level encoder for this purpose. This second encoder reads the embedding vector of each character at each time step, and produces an output vector, and hidden states. Our dual encoder model linearly combines the hidden states of the character-level target expression encoder with the hidden states of the word-level encoder with the following equation:

$$h_{new} = h_1 W_1 + h_2 W_2 + B$$

Here, h_1 is the context representation, and h_2 is the target expression representation. The Bias B and the combination weights W_1 and W_2 are learnable.

4 Experiment

4.1 Dataset

We collect the non-standard English corpus[2] from Urban Dictionary (UD)[3]—the largest online slang dictionary, where terms, definitions and examples are submitted by the crowd. UD is made a reliable resource, due to the quality control of its publishing procedure. To prevent vandalism, a user must have a Facebook or Gmail account, and when each user submits an entry, the UD editors will vote "Publish" or "Don't Publish" (Lloyd, 2011). Each editor is also distinguished by IP addresses, and HTTP cookies are used to prevent each editor from cheating. In recent years, United States Federal government has consulted UD for the definition of "murk" in a threat case (Jackson, 2011), whereas UD is also referred in a financial restitution case in Wisconsin (Kaufman, 2013), as well as determining appropriate license plates in Las Vegas (Davis, 2011).

A total of 421K entries (words and phrases) from the period of 1999-2014 are collected. Each entry includes a list of candidate definitions and examples, as well as the surface form of the target

[2]We have released the dataset for research usage. The dataset is available at: http://www.cs.ucsb.edu/~william/data/slang_ijcnlp.zip

[3]www.urbandictionary.com

Instance: *"dude taht roflcopter was pretty loltastic!!1!"*
Target: loltastic
Reference: something funnyishly fantastic.
Generated Explanation (Single): a really cool , amazing , and good looking .
Generated Explanation (Dual): a word that is extremely awesome .

Instance: *"danny is so jelouse of my work!"*
Target: jelouse
Reference: how unintelligent people who think they are better than someone spells "jealous".
Generated Explanation (Single): your friend ' s way of saying "
Generated Explanation (Dual): a word used to describe a situation , or a person who is a complete idiot .

Instance: *"that sir right there, is being quite adoucheous"*
Target: adoucheous
Reference: a person acting in a conformative manner that causes social upset and violence.
Generated Explanation (Single): when a male is being a male and a male .
Generated Explanation (Dual): the act of being a douchebag .

Figure 3: Some generated explanations from our system.

term. Using the UD API, we can pinpoint the token positions of the words/phrases, and obtain the ground truth labels for tagging.

Our training and testing data use all the examples in an entry (a non-standard expression). We randomly select 371,028 entries for training, resulting in 907,624 sequence pairs of instance and reference explanation. The test set includes 50,000 entries, and 61,330 sentences. Note that all testing target non-standard expressions, instances and examples do not overlap with those in the training dataset.

4.2 Experimental Settings

Our implementation is based on Tensorflow[4]. For input embeddings, we randomly initialize the word vectors. We use stochastic gradient descent with adaptive learning rate to optimize the cross entropy function. We use BLEU scores (Papineni et al., 2002) for the evaluation.

4.3 Quantitative and Qualitative Results

Quantitative experimental results are showed in Table 1. Here we compare the performance of our proposed large dual encoder model to a single context encoder, word-level attentive sequence-to-sequence, as well as a full character-level context encoder model. We use 256 hidden units for this full character-level, because it is the largest setting that fits our Titan X Pascale GPU. Character-level model has longer sequence, which becomes one of its shortages compared with word-level model.

[4]www.tensorflow.org

Model (w. attention)	hidden units	B1	B2
single encoder	1024	21.06	2.1
small dual encoder	512	21.84	2.2
large dual encoder	1024	**24.58**	**2.37**
full char-level model	256	21.13	1.8

Table 1: BLEU scores for explaining non-standard English words and phrases in test dataset.

We see that the single encoder and full character level context models do not offer the best empirical performances on this dataset. Our novel dual encoder method, which combines the strengths of word-level and character-level sequence-to-sequence models, obtained the best performance.

For qualitative analysis, we provide some generated explanations in Figure 3. For example, the first target non-standard expression is the word "loltastic", which is a combination of the words "lol" and "fantastic". To explain this composite non-standard expression, a character-level encoder is needed. The generated explanation from our dual encoder approach clearly makes more sense than the single decoder result.

Overall, dual encoder can explain words with more confidence partly because it knows which words it needs to explain, especially for sentences containing multiple non-standard English words and phrases. Our model can also accurately explain many known acronyms. We also notice that LSTM cells outperform gated recurrent units (GRUs) (Cho et al., 2014), and attention mechanism improves the performance.

5 Conclusion

In this paper, we introduce a new task of learning to explain newly emerged, non-standard English words and phrases. To do this, we collected 15-year of UrbanDictionary data, and designed a dual encoder attentive sequence-to-sequence model to learn the hidden context representation and the hidden non-standard expression embedding. We showed that combining word-level and character-level models improved the performance for this task

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* .

Clint Burfoot and Timothy Baldwin. 2009. Automatic satire detection: Are you having a laugh? In *Proceedings of the ACL-IJCNLP 2009 conference short papers*. Association for Computational Linguistics, pages 161–164.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. *Syntax, Semantics and Structure in Statistical Translation* page 103.

Johnny Davis. 2011. *In praise of urban dictionaries*. The Guardian.

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1277–1287.

Stephan Gouws, Donald Metzler, Congxing Cai, and Eduard Hovy. 2011. Contextual bearing on linguistic variation in social media. In *Proceedings of the Workshop on Languages in Social Media*. Association for Computational Linguistics, pages 20–29.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a# twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pages 368–378.

Leo Hickman. 2013. *Why IBM's Watson supercomputer can't speak slang*. The Guardian.

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

Joe Jackson. 2011. *Feds Check Urban Dictionary to Crack Gun-Store Death Threat*. TIME.

Leslie Kaufman. 2013. *For the Word on the Street, Courts Call Up an Online Witness*. The New York Times.

Leah Kolt and Matt Lazier. 2009. *Alumni in the News: Summer & Fall 2009*. Cal Poly Magazine.

JoAnn Lloyd. 2011. *Alum Aaron Peckham's Urban Dictionary Redefines Language*. Cal Poly Magazine.

Thanapon Noraset, Liang Chen, Larry Birnbaum, and Doug Downey. 2016. Definition modeling: Learning to define word embeddings in natural language. In *AAAI*. pages 3259–3266.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, pages 311–318.

Sara Rosenthal and Kathleen McKeown. 2011. Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pages 763–772.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. pages 3104–3112.

Oriol Vinyals, Łukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. 2015. Grammar as a foreign language. In *Advances in Neural Information Processing Systems*. pages 2773–2781.

William Yang Wang and Kathleen McKeown. 2010. "got you!": Automatic vandalism detection in wikipedia with web-based shallow syntactic-semantic modeling. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*.

Michele Zappavigna. 2012. *Discourse of Twitter and social media: How we use language to create affiliation on the web*. A&C Black.