Unsupervised Neologism Normalization Using Embedding Space Mapping

Nasser Zalmout,[⋆] Aasish Pappu[†] and Kapil Thadani[‡]
[⋆]Computational Approaches to Modeling Language Lab
New York University Abu Dhabi, UAE
[†]Spotify Research, New York, USA
[‡]Yahoo Research, New York, USA
nasser.zalmout@nyu.edu, aasishp@spotify.com, thadani@verizonmedia.com

Abstract

This paper presents an approach for detecting and normalizing neologisms in social media content. Neologisms refer to recent expressions that are specific to certain entities or events and are being increasingly used by the public, but have not yet been accepted in mainstream language. Automated methods for handling neologisms are important for natural language understanding and normalization, especially for informal genres with user generated content. We present an unsupervised approach for detecting neologisms and then normalizing them to canonical words without relying on parallel training data. Our approach builds on the text normalization literature and introduces adaptations to fit the specificities of this task, including phonetic and etymological considerations. We evaluate the proposed techniques on a dataset of Reddit comments, with detected neologisms and corresponding normalizations.

1 Introduction

Linguistic evolution and word coinage are naturally occurring phenomena in languages. However, the proliferation of social media in recent years may expedite these processes by enabling the rapid spread of informal textual content. One aspect of this change is the increasing use of neologisms. Neologisms are relatively recent terms that are used widely and may be in the process of entering common use, but have not yet been fully accepted into mainstream language. Neologisms are rarely found in traditional dictionaries or language lexica, and they usually have lexical, phonetic or semantic connections to some relevant canonical words. They are also often, but not necessarily, generated by combining two different words into a single blend word. Examples include the word *burkini*, which is coined from the words *burka* and *bikini*. The burkini has its own individual meaning that cannot be entailed by a burka or bikini alone.

The goal of neologism normalization is not to generate a perfect replacement for the original text but rather to assist both humans and automated systems in understanding informal text. Inexact normalizations may nevertheless be useful hints to human readers who are unfamiliar with the new words. Normalizations can also substitute for out-of-vocabulary words in downstream NLP applications in order to compensate for data sparsity.

In this paper, we present an unsupervised approach for normalization, based on the hypothesis that neologisms—and non-standard words (NSWs) in general—are likely to share contexts with related canonical words. For instance, NSWs may be expected to lie near their canonical forms in a suitable embedding space. We develop measures to relate words more accurately using both orthography and distributed representations. We also enhance the embedding space with multi-word phrases and subword units, which induces a clustering of compound words with shared etymology, phrases with overlapping words, and entities with common names, thereby capturing novel puns, nicknames, etc.

2 Related Work

Prior work on automatic neologism handling, whether for detection or normalization, is relatively scarce. Most existing neologism detection approaches rely on exclusions lists of canonical or accepted words to filter plausible neologisms (de Yzaguirre, Lluis, 1995; Renouf, 1993). Other contributions based on the same architecture utilize additional filters like eliminating words with spelling errors or named entities to further reduce the set of detected plausible neologisms (Kerremans et al., 2012; Gérard et al., 2014; Cartier, 2016, 2017). There are also several machine learning based approaches, but with limited performance (Falk et al., 2014; Stenetorp, 2010).

In the broader text normalization literature, several supervised approaches have been proposed

Proceedings of the 2019 EMNLP Workshop W-NUT: The 5th Workshop on Noisy User-generated Text, pages 425–430
Hong Kong, Nov 4, 2019.

(Mays et al., 1991; Church and Gale, 1991; Brill and Moore, 2000; Aw et al., 2006; Sproat and Jaitly, 2016), all of which require relatively large datasets. Several unsupervised normalization models have also been presented. Li and Liu (2014); Rangarajan Sridhar (2015) use distributed word embeddings, where the embeddings are used to capture the notion of contextual similarity between canonical and noisy words, along with other measures. Rangarajan Sridhar (2015) further builds on this approach with phrase-based modeling using existing phrase corpora. Hassan and Menezes (2013) use a random-walk based algorithm to calculate contextual similarity, along with edit distance metrics, to obtain normalization candidates. In this paper, we extend the distributed word representation approach (Rangarajan Sridhar, 2015) for unsupervised neologism normalization through several adaptations.

3 Neologism Detection

We first present our neologism and NSW detection approach for Reddit comments. The resulting list of plausible neologisms is then used to analyze neologism etymology and coinage patterns, and later to produce normalization candidates in the normalization model.

Owing to the noisy domain of user-generated text and to the fact that neologisms must exclude names, domain jargon and typos, corpus frequency alone is not reliable for identifying neologisms. Exclusion lists prove effective at recovering a high-precision set of neologisms for this task when combined with frequency-based filters and adaptations to increase coverage. Our pipeline for neologism detection includes the following steps:

- Tokenization: We split on whitespace and handle many Reddit-specific issues, including URLs and specific punctuation patterns.
- Named entity removal: We use the SpaCy NLP toolkit[1] to identify named entities in context and eliminate them from the plausible neologisms list.
- English exclusion lists: We use several corpora of English content as exclusion lists.
- Non-English content removal: We use the Langdetect library[2] to identify and eliminate non-English content.
- Social media jargon removal: We use the social media word clusters from the work by Owoputi et al. (2013) along with the Reddit glossary[3] as additional exclusion lists.

We apply exclusion list filtering on the stem level to further reduce the sparsity of the analysis and reduce the vocabulary. We use NLTK's Snowball stemmer.[4]

4 Neologism Normalization

Our approach is based on the hypothesis that neologisms and NSWs are likely to have similar contexts as their plausible canonical equivalents. We model this using distributed word representations derived from word2vec (Mikolov et al., 2013) via Gensim (Řehůřek and Sojka, 2010). We use these embeddings to learn normalization lexicons and use these lexicons to obtain plausible candidates for normalizing each neologism. We then select among these candidates using a language model and lattice-based Viterbi decoding.

4.1 Lexicon and Lattice Decoding

We use a list of canonical word forms as normalization candidates. This list of canonical forms can be obtained from traditional English language lexica like the Gigaword corpus. For each canonical candidate, we get the N nearest neighbors from the embedding space. This effectively functions as a reversed normalization lexicon, where the canonical candidates are mapped to the potential neologisms. We score the canonical forms using several similarity metrics. We then reverse this mapping to get the list of scored canonical candidates for each neologism.

Neologisms are expected to share semantic, lexical, and phonetic similarity with their canonical counterparts. We capture these different aspects using multiple measures of similarity:

Semantic similarity using the cosine distance over embeddings R_i corresponding to strings S_i.

$$\text{Cos}(S_1, S_2) = \frac{R_1 \cdot R_2}{||R_1|| \times ||R_2||} \quad (1)$$

Lexical similarity based on the formula presented by Hassan and Menezes (2013) and used by Rangarajan Sridhar (2015)

$$\text{Lex}(S_1, S_2) = \frac{LCSR(S_1, S_2)}{ED(S_1, S_2)} \quad (2)$$

[1] Version 2.0.0: https://spacy.io

[2] Version 1.0.7: https://pypi.python.org/pypi/langdetect

[3] https://www.reddit.com/r/TheoryOfReddit/wiki/glossary

[4] Version 3.2.4: http://www.nltk.org/api/nltk.stem.html

where ED is the edit distance and LCSR refers to the longest common subsequence ratio

$$\text{LCSR}(S_1, S_2) = \frac{LCS(S_1, S_2)}{\max(|S_1|, |S_2|)} \quad (3)$$

where LCS is the longest common subsequence in the two strings of length $|S_1|$ and $|S_2|$.

Phonetic similarity through the Metaphone phonetic representation algorithm (Philips, 1990), which is used for indexing words by their English pronunciation. We calculate the normalized edit distance for the Metaphone representation of S_1 and S_2 and use this score to reflect the phonetic similarity between the strings.

$$\text{PHON}(S_1, S_2) = 1 - \frac{ED(mP(S_1), mP(S_2))}{\max(|S_1|, |S_2|)} \quad (4)$$

where $mP(S_i)$ is a Metaphone representation.

Next, a language model is used to further control the fluency of the normalized output in context. We use SRILM (Stolcke, 2002) to build the model. To decode the optimal path given the similarity scores and the language model probabilities, we encode the sentence, along with the various normalization candidates, in the HTK format. We then use SRILM's lattice-tool toolkit to decode the space of potential paths using Viterbi decoding.

4.2 Phrases and Subword Units

The system so far is primarily targeted to word-level normalization, without explicitly handling multi-word phrases in the canonical form or recognizing shared etymology in the embeddings for plausible neologisms. This limits the normalization space for neologisms as the blending of two or more words is a common neologism pattern.

Multi-word phrases: We use a data-driven approach for identifying common phrases within the given corpus (Mikolov et al., 2013). Phrase candidates with scores above a certain threshold have their constituent words joined by a delimiter and are considered as a single word-like token for subsequent analysis. We ensure that the detected phrases do not contain punctuation sequences or URLs, which are common in Reddit data.

Subword units: Traditional morphology-based analysis falls short of detecting proper subword entities in neologisms, where the form and etymology of the neologisms are not fixed. Moreover, n-gram character sequences are also not optimal here given the intractability of the analysis. Instead we segment words based on the *byte pair encoding* (BPE) algorithm (Gage, 1994), which was adapted for use in neural machine translation (NMT) by Sennrich et al. (2016).

We add the detected phrases to the list of canonical candidates as potential normalization targets and add the subwords to the neologism lists.

4.3 Combining Word Representations

An important aspect to consider when combining word, phrase and subword representations is to maintain the distributional properties of the text. We combine these representations by having the choice to switch to a certain representation for each word dictated through a uniformly distributed random variable. That is, for a given sentence T in a corpus, and for each word $w_i \in T$, the resulting representation w_i' based on the distribution $q(w_i'|w_i)$ is managed by the control variable $c = rand(\alpha)$, where $\alpha \in \{0, 1, 2\}$ indicates the choice of word/phrase/subword representations. We repeat this process for all the words of each sentence k different times, so we end up with k different copies of the sentence, each having a randomly selected representation for all of its words. k is tunable and we set $k = 5$ for our experiments. A somewhat similar approach is used by Wick et al. (2016) to learn multilingual word embeddings.

5 Experimental Setup and Results

5.1 Dataset

We use a dataset of Reddit comments from June 2016 to June 2017 for the normalization experiments in this paper, collected with the Reddit BigQuery API.[5] We focus on five popular subreddit groups: *worldnews, news, politics, sports,* and *movies*. This dataset contains about 51M comments, 2B tokens (words), and 6M unique words.

A dataset of 2034 comments annotated with neologisms and their normalizations was used for tuning[6] and evaluating the normalization model. These comments were selected from comments identified as containing unique plausible neologisms using the neologism detection pipeline described in Section 3. Normalization annotations were obtained using Amazon Mechanical Turk using three judgments per comment. Annotators were asked to provide up to five normalization candidates for each neologism; candidates that at

[5] https://bigquery.cloud.google.com/dataset/fh-bigquery:reddit_comments

[6] Parameters were tuned using a held-out validation set drawn from the manual neologism annotations. This also applies to the tuning of weights for the linear combination of the different similarity metrics.

	Sample of detected neologisms
politics	pizzagate, drumpf, trumpster, shillary, killary
news	antifa, brexit, drumpf, Libruls, redpilling, neonazi
worldnews	burkini, brexit, pizzagate, edgelord, petrodollar
sports	deflategate, handegg, ballboy, skurfing, playstyle
movies	plothole, stuckmannized, jumpscare, MetaHuman
gaming	playerbase, pokestop, jumpscare, hitscan

Table 1: Subreddit-level detected neologisms

	Accuracy	BLEU
Baseline	55.3	81.3
This work	64.2	87.7

Table 2: Evaluation of the normalization systems

least two of the three annotators agree upon were selected as normalizations.[7]

5.2 Neologism Detection

We apply the detection pipeline we discussed earlier on the Reddit dataset. We use the most frequent 64K words in the Gigaword corpus as an exclusion list for proper English words along with NLTK's Words and Wordnet corpora. We further eliminated the terms that had a frequency lower than 10 as potential spelling errors.

Table 1 shows samples of the top detected neologisms for each subreddit. We took a random sample of 500 Reddit comments to inspect manually. Based on our observations, 5% of the comments contained neologisms, and 82% of these neologisms are present in our list of plausible neologisms, which suggests the recall of the proposed detection pipeline.

5.3 Normalization

We trained the word2vec model using the Reddit dataset using the skip-gram algorithm, a window of 5 words, and an embedding size of 250. For phrase learning, we used a threshold score of 10 and minimum count of 5. For lattice decoding, we used a trigram language model with Kneser-Ney discounting trained on LDC's Gigaword Fourth Edition corpus (LDC2009T13) (Parker et al., 2009).

As a baseline, we use the model of Rangarajan Sridhar (2015), which does not consider subwords and phonological similarity. They use language models and lattices, similar to our work, but targeted for text normalization. Our work extends these ideas to normalize a wide variety of neologisms including phrases, nicknames and compound words.

For evaluation metrics, we use the accuracy of the normalization on the word level (the neologisms/canonical-equivalents level) along with using BLEU score (Papineni et al., 2002). BLEU is an algorithm for evaluating text quality based on human references and is commonly used in the machine translation literature. Using BLEU is relevant here due to the potentially multi-word output of the system with phrases and subwords. Evaluation scores are calculated with some relaxed matching, namely considering the occurrences of plurals, lower/upper case, hyphenation and punctuation, among others. So we treat terms like *trump* and *Trumps* as equivalent, same for *posting* and *postings*.

[7] We started with a dataset of 5000 unique neologisms and eliminated those that did not have a consensus of two or that the annotators indicated they were not sure about.

	Sentence
Raw	republicans who don't want drumpf are voting for hilldawg
Best system	republicans who don't want trump are voting for hillary
Reference	republicans who don't want trump are voting for hillary
Raw	this is one of the biggest clickbate news outlets
Best system	this is one of the biggest click bait news outlets
Reference	this is one of the biggest click bait news outlets
Raw	the hillbots have gone full insanity
Best system	the hill bots have gone full insanity
Reference	the hillary bots have gone full insanity

Table 3: Normalization examples

Table 2 shows the results. The system with phrases and subwords clearly outperforms the baseline, for both accuracy and BLEU scores. BLEU scores are relatively high for both systems since most of the sentences are preserved with only modifications for the plausible neologisms. The rest of the sentence should be an exact match to the reference.

Table 3 presents three normalization examples, with the raw, gold reference, and the output of our system. The examples show a promising behavior, but as can be seen at the third example, there is still a room for improvement in normalizing the individual phrase components. A potential future direction here would be to improve embedding space mappings for the subword entities.

6 Conclusion

We presented an approach to detect and normalize neologisms in social media content. We leveraged the fact that the neologisms and their canonical equivalents are likely to share the same contexts and hence have relatively close distributional representations. We also presented some techniques for handling phrases and subwords in the plausible neologisms, which is important given the etymology behind most neologisms. Our approach also makes use of the phonetic representation of the words, to capture coinage patterns that involve phonetic-based modification. Our results show that the model is effective in both detection and

normalization. Future work includes more explicit generation models, utilizing natural language generation techniques, along with expanding and enhancing the coverage of the annotated data.

References

AiTi Aw, Min Zhang, Juan Xiao, and Jian Su. 2006. A phrase-based statistical model for sms text normalization. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, COLING-ACL '06, pages 33–40, Stroudsburg, PA, USA. Association for Computational Linguistics.

Eric Brill and Robert C. Moore. 2000. An improved error model for noisy channel spelling correction. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, pages 286–293, Stroudsburg, PA, USA. Association for Computational Linguistics.

Emmanuel Cartier. 2016. Neoveille, système de repérage et de suivi des néologismes en sept langues. *Neologica : revue internationale de la néologie*, (10).

Emmanuel Cartier. 2017. Neoveille, a web platform for neologism tracking. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 95–98, Valencia, Spain. Association for Computational Linguistics.

Kenneth W. Church and William A. Gale. 1991. Probability scoring for spelling correction. *Statistics and Computing*, 1(2):93–103.

CabrÃl', Maria ; de Yzaguirre, Lluis. 1995. StratÃl'gie pour la dÃl'tection semi-automatique des nÃl'ologismes de presse. *TTR*, 8(2):89–100.

Ingrid Falk, Delphine Bernhard, and Christophe Gérard. 2014. From non word to new word: Automatically identifying neologisms in french newspapers. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA). ACL Anthology Identifier: L14-1260.

Philip Gage. 1994. A new algorithm for data compression. *C Users J.*, 12(2):23–38.

Christophe Gérard, Ingrid Falk, and Delphine Bernhard. 2014. Traitement automatisé de la néologie: pourquoi et comment intégrer l'analyse thématique? In *SHS Web of Conferences*, volume 8, pages 2627–2646. EDP Sciences.

Hany Hassan and Arul Menezes. 2013. Social text normalization using contextual graph random walks. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1577–1586, Sofia, Bulgaria. Association for Computational Linguistics.

Daphné Kerremans, Susanne Stegmayr, and Hans-Jörg Schmid. 2012. The neocrawler: identifying and retrieving neologisms from the internet and monitoring on-going change. *Current methods in historical semantics*, 73:59.

Chen Li and Yang Liu. 2014. Improving text normalization via unsupervised model and discriminative reranking. In *Proceedings of the ACL 2014 Student Research Workshop*, pages 86–93, Baltimore, Maryland, USA. Association for Computational Linguistics.

Eric Mays, Fred J. Damerau, and Robert L. Mercer. 1991. Context based spelling correction. *Information Processing & Management*, 27(5):517 – 522.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–390, Atlanta, Georgia. Association for Computational Linguistics.

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*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2009. English gigaword fourth edition ldc2009t13. *Linguistic Data Consortium, Philadelphia*.

Lawrence Philips. 1990. Hanging on the Metaphone. *Computer Language*, 7(12 (December)).

Vivek Kumar Rangarajan Sridhar. 2015. Unsupervised text normalization using distributed representations of words and phrases. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 8–16, Denver, Colorado. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Antoinette Renouf. 1993. Sticking to the text: a corpus linguistâĂŹs view of language. *ASLIB*, 45(5):131–136.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Richard Sproat and Navdeep Jaitly. 2016. RNN approaches to text normalization: A challenge. *CoRR*, abs/1611.00068.

Pontus Stenetorp. 2010. *Automated extraction of swedish neologisms using a temporally annotated corpus*. MasterâĂŹs thesis, Royal Institute of Technology(KTH), Stockholm, Sweden.

Andreas Stolcke. 2002. Srilm – an extensible language modeling toolkit. In *In Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002*, pages 901–904.

Michael Wick, Pallika Kanani, and Adam Pocock. 2016. Minimally-constrained multilingual embeddings via artificial code-switching. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 2849–2855. AAAI Press.