Break it Down for Me: A Study in Automated Lyric Annotation

Lucas Sterckx[*], **Jason Naradowsky**[†], **Bill Byrne**[‡], **Thomas Demeester**[*] and **Chris Develder**[*]
[*] IDLab, Ghent University - imec
firstname.lastname@ugent.be
[†] Language Technology Lab, DTAL, University of Cambridge
jrn39@cam.ac.uk
[‡] Department of Engineering, University of Cambridge
wjb31@cam.ac.uk

Abstract

Comprehending lyrics, as found in songs and poems, can pose a challenge to human and machine readers alike. This motivates the need for systems that can understand the ambiguity and jargon found in such creative texts, and provide commentary to aid readers in reaching the correct interpretation.

We introduce the task of automated lyric annotation (ALA). Like text simplification, a goal of ALA is to rephrase the original text in a more easily understandable manner. However, in ALA the system must often include *additional* information to clarify niche terminology and abstract concepts. To stimulate research on this task, we release a large collection of crowdsourced annotations for song lyrics. We analyze the performance of translation and retrieval models on this task, measuring performance with both automated and human evaluation. We find that each model captures a unique type of information important to the task.

1 Introduction

Song lyrics and poetry often make use of ambiguity, symbolism, irony, and other stylistic elements to evoke emotive responses. These characteristics sometimes make it challenging to interpret obscure lyrics, especially for readers or listeners who are unfamiliar with the genre. To address this problem, several online lyric databases have been created where users can explain, contextualize, or discuss lyrics. Examples include MetroLyrics[1] and Genius.com[2]. We refer to such commentary as a lyric annotation (Figure 1).

How does it feel?
To be without a home
Like a complete unknown,
Like a rolling stone

↓

The proverb "A rolling stone gathers no moss" refers to people who are always on the move, never putting down roots or accumulating responsibilities and cares.

Figure 1: A lyric annotation for "Like A Rolling Stone" by Bob Dylan.

In this work we introduce the task of *automated lyric annotation* (ALA). Compared to many traditional NLP systems, which are trained on newswire or similar text, an automated system capable of explaining abstract language, or finding alternative text expressions for slang (and other unknown terms) would exhibit a deeper understanding of the nuances of language. As a result, research in this area may open the door to a variety of interesting use cases. In addition to providing lyric annotations, such systems can lead to improved NLP analysis of informal text (blogs, social media, novels and other literary works of fiction), better handling of genres with heavy use of jargon (scientific texts, product manuals), and increased robustness to textual variety in more traditional NLP tasks and genres.

Our contributions are as follows:

1. To aid in the study of ALA we present a corpus of 803,720 crowdsourced lyric annotation pairs suitable for training models for this task.[3]
2. We present baseline systems using statistical machine translation (SMT), neural trans-

[1] http://www.metrolyrics.com
[2] http://genius.com

[3] To obtain the data collection please contact the first author of this paper.

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2063–2069
Copenhagen, Denmark, September 7–11, 2017.

# Lyric Annotation pairs	803,720
⊘ Tokens per Lyric	15
⊘ Tokens per Annotation	43
$\|V_{\text{lyrics}}\|$	124,022
$\|V_{\text{annot}}\|$	260,427

Table 1: Properties of gathered dataset (V_{lyrics} and V_{annot} denote the vocabulary for lyrics and annotations, ⊘ denotes the average amount).

lation (Seq2Seq), and information retrieval.

3. We establish an evaluation procedure which adopts measures from machine translation, paraphrase generation, and text simplification. Evaluation is conducted using both human and automated means, which we perform and report across all baselines.

2 The Genius ALA Dataset

We collect a dataset of crowdsourced annotations, generated by users of the *Genius* online lyric database. For a given song, users can navigate to a particular stanza or line, view existing annotations for the target lyric, or provide their own annotation. Discussion between users acts to improve annotation quality, as it does with other collaborative online databases like Wikipedia. This process is gamified: users earn *IQ* points for producing high quality annotations.

We collect 736,423 lyrics having a total 1,404,107 lyric annotation pairs from all subsections (rap, poetry, news, etc.) of Genius. We limit the initial release of the annotation data to be English-only, and filter out non-English annotations using a pre-trained language identifier. We also remove annotations which are solely links to external resources, and do not provide useful textual annotations. This reduces the dataset to 803,720 lyric annotation pairs. We list several properties of the collected dataset in Table 1.

2.1 Context Independent Annotation

Mining annotations from a collaborative human-curated website presents additional challenges worth noting. For instance, while we are able to generate large quantities of parallel text from Genius, users operate without a single, predefined and shared *global* goal other than to maximize their own IQ points. As such, there is no motivation to provide annotations for a song in its entirety, or independent of previous annotations.

For this reason we distinguish between two types of annotations: *context independent* (CI) annotations are independent of their surrounding context and can be interpreted without it, e.g., explain specific metaphors or imagery or provide narrative while normalizing slang language. Contrastively, *context sensitive* (CS) annotations provide broader context beyond the song lyric excerpt, e.g., background information on the artist.

To estimate contribution from both types to the dataset, we sample 2,000 lyric annotation pairs and label them as either CI or CS. Based on this sample, an estimated 34.8% of all annotations is independent of context. Table 2 shows examples of both types.

While the goal of ALA is to generate annotations of all types, it is evident from our analysis that CS annotations can not be generated by models trained solely on parallel text. That is, these annotations cannot be generated without background knowledge or added context. Therefore, in this preliminary work we focus on predicting CI lyric annotations.

3 Baselines

We experiment with three baseline models used for text simplification and paraphrase generation.

- **Statistical Machine Translation (SMT):** One approach is to treat the task as one of translation, and to use established statistical machine translation (SMT) methods (Quirk et al., 2004) to produce them. We train a standard phrase-based SMT model to translate lyrics to annotations, using GIZA++ (Josef Och and Ney, 2003) for word alignment and Moses (Koehn et al., 2007) for phrasal alignment, training, and decoding.

- **Seq2Seq:** Sequence-to-sequence models (Sutskever et al., 2014) offer an alternative to SMT systems, and have been applied successfully to a variety of tasks including machine translation. In Seq2Seq, a recurrent neural network (RNN) encodes the source sequence to a single vector representation. A separate decoder RNN generates the translation conditioned on this representation of the source sequence's semantics. We utilize Seq2Seq with attention (Bahdanau et al., 2014), which allows the model to

Type	% of annotations	Examples	
CI (Context independent)	34.8%	[L]	Gotta patch a lil kid tryna get at this cabbage
		[A]	He's trying to ignore the people trying to get at his money.
		[L]	You know it's beef when a smart brother gets stupid
		[A]	You know an argument is serious when an otherwise rational man loses rational.
CS (Context sensitive)	65.2%	[L]	Cause we ain't break up, more like broke down
		[A]	The song details Joe's break up with former girlfriend Esther.
		[L]	If I quit this season, I still be the greatest, funk
		[A]	Kendrick has dropped two classic albums and pushed the artistic envelope further.

Table 2: Examples of context independent and dependent pairs of lyrics [L] and annotations [A].

additionally condition on tokens from the input sequence during decoding.

- **Retrieval:** In practice, similar lyrics may reappear in different contexts with exchangeable annotations. We treat the training corpus as a database of lyrics' excerpts with corresponding annotations, and at test time select the annotation assigned to the most similar lyric. This baseline is referred to as the *retrieval* model. We use standard TF-IDF weighted cosine distance as similarity measure between lyrics' excerpts.

4 Evaluation

4.1 Data

We evaluate automatic annotators on a selection of 354 CI annotations and partition the rest of the annotations into 2,000 instances for development and the full remainder for training. It is important to note that the annotations used for training and development include CI as well as CS annotations.

Annotations often include multiple sentences or even paragraphs for a single lyrics excerpt (which does not include end marks), while machine translation models need aligned corpora at sentence level to perform well (Xu et al., 2016). We therefore transform training data by including each sentence from the annotation as a single training instance with the same lyric, resulting in a total of 1,813,350 sentence pairs.

We use this collection of sentence pairs (denoted as *sent.* in results) to train the SMT model. Seq2Seq models are trained using sentence pairs as well as full annotations. Interestingly, techniques encouraging alignment by matching length and thresholding cosine distance between lyric and annotation did not improve performance during development.

4.2 Measures

For automated evaluation, we use measures commonly used to evaluate translation systems (BLEU, METEOR), paraphrase generation (iBLEU) and text simplification (SARI).

BLEU (Papineni et al., 2002) uses a modified form of precision to compare generated annotations against references from Genius. METEOR (Denkowski and Lavie, 2011) is based on the harmonic mean of precision and recall and, along with exact word matching, includes stemming and synonymy matching. iBLEU (Sun and Zhou, 2012) is an extension of the BLEU metric to measure diversity as well as adequacy of the annotation, iBLEU $= 0.9 \times$ BLEU(Annotation, Reference) $- 0.1 \times$ BLEU(Annotation, Lyric). SARI (Xu et al., 2016) measures precision and recall of words that are added, kept, or deleted separately and averages their arithmetic means.

We also measure quality by crowdsourcing ratings via the online platform CrowdFlower.[4] We present collaborators with a song lyric excerpt annotated with output from the annotation generators as well as a reference annotation from Genius. Collaborators assign a 5-point rating for *Fluency* which rates the quality of the generated language, and *Information* which measures the added clarification by the annotation, a key aspect of this task. For each lyric annotation pair, we gather ratings from three different collaborators and take the average.

4.3 Hyperparameters and Optimization

Here we describe implementation and some of the optimizations used when training the models.

[4]https://www.crowdflower.com/

Lyric:	`and when I slept, I dream g's, son I need some`
Human:	he dreams about large amounts of money because he is in need of it .
SMT:	and when he slept, and he's dream, g's son, he needs some
Seq2Seq:	when he sleeps , he dreams of getting money.
Retrieval:	g's refer to the ghetto gangsters in london that the kid has become involved with
Lyric:	`Mossberg pump I'm ridin' shotgun literally`
Human:	riding shotgun means to ride up front in the passenger seat
SMT:	Mossberg pump - action shotgun, he's saying he's ridin' shotgun literally.
Seq2Seq:	Mossberg is a firearms manufacturer. He is riding shotgun with his shotgun.
Retrieval:	he fully respects the people mentioned earlier and is riding for them.
Lyric:	`and my eyes red cause of all that haze`
Human:	his eyes turn red due to lack of hydration from the weed one smokes (in this case , smoking purple haze)
SMT:	and his eyes are red because of all that haze.
Seq2Seq:	his eyes are red because of all the weed he smokes.
Retrieval:	his eyes are red as a sign of the fatigue of travel and , possibly , from drugs and alcohol .

Table 3: Lyrics excerpts with annotations from Genius ('Human') and automated annotators.

	Properties		Automated Evaluation				Human Evaluation	
	Length Ratio	Profanity/Tok.	BLEU	iBLEU	METEOR	SARI	Fluency	Information
Human	1.19	0.0027	-	-	-	-	3.93	3.53
SMT (Sent.)	1.23	0.0068	6.22	1.44	12.20	38.42	3.82	3.31
Seq2Seq (Sent.)	1.05	0.0023	5.33	3.64	9.28	36.52	3.76	3.25
Seq2Seq	1.32	0.0022	5.15	3.46	10.56	36.86	3.83	3.34
Retrieval	1.18	0.0038	2.82	2.27	5.10	32.76	3.93	2.98

Table 4: Quantitative evaluation of different automated annotators.

For Seq2Seq models, we use OpenNMT (Klein et al., 2017) and optimize for perplexity on the development set. Vocabulary for both lyrics and annotations is reduced to the 50,000 most frequent tokens and are embedded in a 500-dimensional space.

We use two layers of stacked bi-directional LSTMs with hidden states of 1024 dimensions. We regularize using dropout (keep probability of 0.7) and train using stochastic gradient descent with batches of 64 samples for 13 epochs.

The decoder of the SMT model is tuned for optimal BLEU scores on the development set using minimum error rate training (Bertoldi et al., 2009).

5 Results

To measure agreement between collaborators, we compute the kappa statistic (Fleiss, 1971). Kappa statistics for fluency and information are 0.05 and 0.07 respectively, which indicates low agreement. The task of evaluating lyric annotations was difficult for CrowdFlower collaborators as was apparent from their evaluation of the task. For evaluation in future work, we recommend recruitment of expert collaborators familiar with the Genius platform and song lyrics.

Table 3 shows examples of lyrics with annotations from Genius and those generated by baseline models.

A notable observation is that translation models learn to take the role of narrator, as is common in CI annotations, and recognize slang language while simplifying it to more standard English.

Automatic and human evaluation scores are shown in Table 4. Next to evaluation metrics, we show two properties of automatically generated annotations; the average annotation length relative to the lyric and the occurrence of profanity per token in annotations, using a list of 343 swear words.

The SMT model scores high on BLEU, METEOR and SARI but shows a large drop in performance for iBLEU, which penalizes lexical similarity between lyrics and generated annotations as apparent from the amount profanity remaining in the generated annotations.

Standard SMT rephrases the song lyric from a third person perspective but is conservative in lexical substitutions and keeps close to the grammar of the lyric. A more appropriate objective function for tuning the decoder which promotes lexical dissimilarity as done for paraphrase generation, would be beneficial for this approach.

Seq2Seq models generate annotations more dissimilar to the song lyric and obtain higher iBLEU

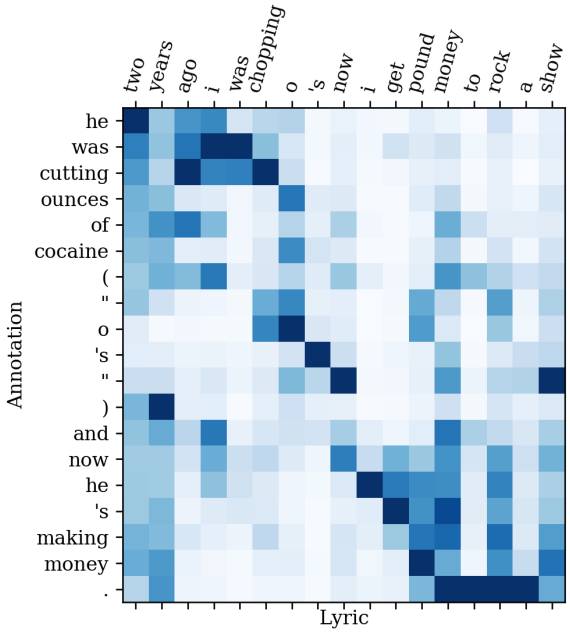

Figure 2: Attention visualization of Seq2Seq models for ALA.

and Information scores. To visualize some of the alignments learned by the translation models, Fig. 2 shows word-by-word attention scores for a translation by the Seq2Seq model.

While the retrieval model obtains quality annotations when test lyrics are highly similar to lyrics from the training set, retrieved annotations are often unrelated to the test lyric or specific to the song lyric it is retrieved from.

Out of the unsupervised metrics, METEOR obtained the highest Pearson correlation (Pearson, 1895) with human ratings for Information with a coefficient of 0.15.

6 Related Work

Work on modeling of social annotations has mainly focused on the use of topic models (Iwata et al., 2009; Das et al., 2014) in which annotations are assumed to originate from topics. They can be used as a preprocessing step in machine learning tasks such as text classification and image recognition but do not generate language as required in our ALA task.

Text simplification and paraphrase generation have been widely studied. Recent work has highlighted the need for large text collections (Xu et al., 2015) as well as more appropriate evaluation measures (Xu et al., 2016; Galley et al., 2015). They indicated that especially informal language, with its high degree of lexical variation, e.g., as used in social media or lyrics, poses serious challenges (Xu et al., 2013).

Text generation for artistic purposes, such as poetry and lyrics, has been explored most commonly using templates and constraints (Barbieri et al., 2012). In regard to rap lyrics, Wu et al. (2013) present a system for rap lyric generation that produces a single line of lyrics that is meant to be a response to a single line of input. Most recent work is that of Zhang et al. (2014) and Potash et al. (2015), who show the effectiveness of RNNs for the generation of poetry and lyrics.

The task of annotating song lyrics is also related to metaphor processing. As annotators often explain metaphors used in song lyrics, the Genius dataset can serve as a resource to study computational modeling of metaphors (Shutova and Teufel, 2010).

7 Conclusion and Future Work

We presented and released the Genius dataset to study the task of Automated Lyric Annotation. As a first investigation, we studied automatic generation of context independent annotations as machine translation and information retrieval. Our baseline system tests indicate that our corpus is suitable to train machine translation systems.

Standard SMT models are capable of rephrasing and simplifying song lyrics but tend to keep close to the structure of the song lyric. Seq2Seq models demonstrated potential to generate more fluent and informative text, dissimilar to the lyric.

A large fraction of the annotations is heavily based on context and background knowledge (CS), one of their most appealing aspects. As future work we suggest injection of structured and unstructured external knowledge (Ahn et al., 2016) and explicit modeling of references (Yang et al., 2016).

Acknowledgments

The authors wish to thank the anonymous reviewers for their helpful comments. This work was supported by the Research Foundation - Flanders (FWO) and the U.K. Engineering and Physical Sciences Research Council (EPSRC grant EP/L027623/1).

References

S. Ahn, H. Choi, T. Pärnamaa, and Y. Bengio. 2016. A Neural Knowledge Language Model. *ArXiv e-prints* https://arxiv.org/abs/1608.00318.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR* abs/1409.0473. http://arxiv.org/abs/1409.0473.

Gabriele Barbieri, François Pachet, Pierre Roy, and Mirko Degli Esposti. 2012. Markov constraints for generating lyrics with style. In *ECAI 2012 - 20th European Conference on Artificial Intelligence. Including Prestigious Applications of Artificial Intelligence (PAIS-2012) System Demonstrations Track, Montpellier, France, August 27-31 , 2012*. pages 115–120. https://doi.org/10.3233/978-1-61499-098-7-115.

Nicola Bertoldi, Barry Haddow, and Jean-Baptiste Fouet. 2009. Improved minimum error rate training in moses. *Prague Bull. Math. Linguistics* 91:7–16. http://ufal.mff.cuni.cz/pbml/91/art-bertoldi.pdf.

Mrinal Kanti Das, Trapit Bansal, and Chiranjib Bhattacharyya. 2014. Going beyond corr-lda for detecting specific comments on news & blogs. In *Seventh ACM International Conference on Web Search and Data Mining, WSDM 2014, New York, NY, USA, February 24-28, 2014*. pages 483–492. https://doi.org/10.1145/2556195.2556231.

Michael Denkowski and Alon Lavie. 2011. Proceedings of the sixth workshop on statistical machine translation pages 85–91. http://aclweb.org/anthology/W11-2107.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76(5):378.

Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. 2015. deltableu: A discriminative metric for generation tasks with intrinsically diverse targets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, pages 445–450. https://doi.org/10.3115/v1/P15-2073.

Tomoharu Iwata, Takeshi Yamada, and Naonori Ueda. 2009. Modeling social annotation data with content relevance using a topic model. In *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada.*. pages 835–843. http://papers.nips.cc/paper/3773-modeling-social-annotation-data-with-content-relevance-using-a-topic-model.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics, Volume 29, Number 1, March 2003* http://aclweb.org/anthology/J03-1002.

G. Klein, Y. Kim, Y. Deng, J. Senellart, and A.M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints* https://arxiv.org/abs/1701.02810.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Association for Computational Linguistics, pages 177–180. http://aclweb.org/anthology/P07-2045.

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 of the Association for Computational Linguistics*. http://aclweb.org/anthology/P02-1040.

Karl Pearson. 1895. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London* 58:240–242.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2015. Ghostwriter: Using an lstm for automatic rap lyric generation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1919–1924. https://doi.org/10.18653/v1/D15-1221.

Chris Quirk, Chris Brockett, and William Dolan. 2004. Proceedings of the 2004 conference on empirical methods in natural language processing. http://aclweb.org/anthology/W04-3219.

Ekaterina Shutova and Simone Teufel. 2010. Metaphor corpus annotated for source - target domain mappings. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. European Languages Resources Association (ELRA). http://aclweb.org/anthology/L10-1419.

Hong Sun and Ming Zhou. 2012. Joint learning of a dual smt system for paraphrase generation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, pages 38–42. http://aclweb.org/anthology/P12-2008.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural net-

works. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. pages 3104–3112. http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.

Dekai Wu, Karteek Addanki, Markus Saers, and Meriem Beloucif. 2013. Learning to freestyle: Hip hop challenge-response induction via transduction rule segmentation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 102–112. http://aclweb.org/anthology/D13-1011.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association of Computational Linguistics* 3:283–297. http://aclweb.org/anthology/Q15-1021.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association of Computational Linguistics* 4:401–415. http://aclweb.org/anthology/Q16-1029.

Wei Xu, Alan Ritter, and Ralph Grishman. 2013. Proceedings of the sixth workshop on building and using comparable corpora. Association for Computational Linguistics, pages 121–128. http://aclweb.org/anthology/W13-2515.

Z. Yang, P. Blunsom, C. Dyer, and W. Ling. 2016. Reference-Aware Language Models. *ArXiv e-prints* https://arxiv.org/abs/1611.01628.

Xingxing Zhang and Mirella Lapata. 2014. Chinese poetry generation with recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pages 670–680. https://doi.org/10.3115/v1/D14-1074.