MusicStand: Listening to Song Lyrics Using a Map Query Interface

Ai-Te Kuo
aitekuo@auburn.edu
Auburn University
Auburn, Alabama, USA

Hanan Samet
hjs@cs.umd.edu
University of Maryland
College Park, Maryland, USA

ABSTRACT

Music is present in numerous forms in our daily lives and is deemed essential to it. Multiple applications have been proposed to let users check into a location and tag that check-in with the song to which they are listening. This is time-consuming and requires much work in voluntary manual tagging. One of our major goals is to automatically determine the spatial scope of a song. Our research challenge is how to identify locations in unstructured and badly-cased lyric texts (e.g., all caps, camel case, non-cased, studly caps, etc.) that are mostly submitted by volunteers from all over the world. Uncertain casing leads to a severe performance drop when using named entity recognition (NER) and geographical information is often lost due to a failure to correctly identify geographical entities. We overcome this failure by normalizing the lyrics in the sense that the information loss is minimized and propose the MusicStand(`http://musicstand.umiacs.io/`) framework to process/input lyric text that involves three steps: cleaning, truecasing, and geotagging. MusicStand enables users to explore or search a music collection where the goal is to find and play songs about particular geographic entities (i.e., toponyms) using a map query interface. Note that the collection may be static (e.g., a songbook) or dynamic (e.g., a radio playlist).

CCS CONCEPTS

• **Information systems** → **Information retrieval**.

KEYWORDS

map query interface, geographical information retrieval, geotagging, truecasing

ACM Reference Format:
Ai-Te Kuo and Hanan Samet. 2021. MusicStand: Listening to Song Lyrics Using a Map Query Interface. In *29th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '21), November 2–5, 2021, Beijing, China.* ACM, New York, NY, USA, 4 pages. `https://doi.org/10.1145/3474717.3484211`

SIGSPATIAL '21, November 2–5, 2021, Beijing, China

ACM ISBN 978-1-4503-8664-7/21/11…$15.00
`https://doi.org/10.1145/3474717.3484211`

1 INTRODUCTION

Music is ubiquitous in our daily lives and can provoke a strong sense of emotional resonance. Songs have the ability to simultaneously connect us to the good old days and help us understand the world around us.

Current audio streaming platforms such as Apple Music, Spotify, YouTube Music, and Pandora Radio let you browse songs by artist, genre, or geography. However, these systems only present the most popular songs in a certain geographical area but not the songs that mention the area. Unfortunately, users cannot search songs related to the places in which they are interested. Additionally, current platforms do not permit users to see which keywords are used most frequently in songs pertaining to a certain geographical area. Driven by this problem, we aim to provide users with an intuitive interface that enables them to explore a music collection and a visual representation of mentions about certain places by the use of panning and zooming.

When creating this system, one of the primary challenges comes from obtaining a proper lyric dataset due to copyright restrictions. Lyric datasets are mostly inaccessible or appear in an undesired form to avoid a lawsuit. For example, Million Song Dataset[1] provides a collection of song features and metadata for one million popular songs. Unfortunately, the dataset does not include whole lyrics but appears in a bag-of-words format with 5000 most frequently occurring words. However, over 25 million geographical names have been recorded according to GeoNames [34], and, thus, most locations will not appear in the vocabulary of the most frequently used words. Therefore, we implemented a scraper that collects complete lyric data from different sources and a cleaner that transforms raw HTML text to clean plain text. Obtaining accurate lyrics is of critical importance to help us identify all geographic entities associated with the song.

In many case-sensitive languages, capitalization plays an important role in determining the meaning of the word and the part of speech. For example, consider the following badly-cased sentence "I went to la". It is obvious **la** in this context is actually referring to the city of Los Angeles, but the machine concludes a high probability of **la** being a verb rather than a proper noun. Another example from "House of the Rising Sun" as performed by The Animals is presented in Figure 1, where the upper paragraph contains case information, and the lower paragraph does not. The words in bold are identified by the named-entity recognition (NER) system. We see that none of the named entities from the non-cased text are recognized by the NER system. This occurs because the use of capitalization in the English language signifies a word's importance, especially when referring to geographic locations. Additionally, most NER systems

[1] http://millionsongdataset.com

are trained on well-edited text and are less equipped to handle the unexpected capitalization of words. Unfortunately, lyrics cannot be processed in the same way as news articles, which are typically published after careful review. Because songs are transcribed by anonymous volunteers at several timestamps on popular lyric websites such as AZLyrics [2], Genius [7], and Musixmatch [18], lyrics often come without capitalization information or are inconsistent in their capitalization of words. Therefore, it is challenging to recognize what words refer to a place without a case indicator. To overcome this issue, we follow the primary strategy used by most of the literature [3, 14, 19, 30], which restores the true case of words by using a truecaser before feeding into NER systems.

There is a house in **New Orleans (GPE)**
They call **the Rising Sun (LOC)**
And it's been the ruin of many a poor boy
And God, I know I'm one

there is a house in new orleans
they call the rising sun
and it's been the ruin of many a poor boy
and God, i know i'm one

Figure 1: NER performance degradation in non-cased lyric.

One of the research challenges is to determine the geographical coordinates of a given lyric (known as geotagging). Lyrics are nothing like news articles and can contain badly hyphenated words, fictitious places, grammatical mistakes, jargon terms, made-up words, phonetic spelling, e.g., A-T-L-A-N-T-A-G-A, New York Cityyy, Califor-Nye-Aye, America-Ca-Ca. All the above-mentioned word usages raise difficulties in recognizing toponyms and resolving ambiguity. In addition, some songwriters might prefer using a nickname to refer to a place that is not in the gazetteer data. Some of them are easy to figure out as they are in a global lexicon, but some jargon terms can be difficult to identify as they only can be inferred from inhabitants of the surrounding places. Figure 2 presents the nicknames or the jargon that have been used in songs.

Atlanta (A-town), Boston (The Hub or Beantown)
Chicago (Chi-town or The Windy City)
Dallas (Big D or D town), Houston (H-town)
The Carolinas (Cackalacky or Cackalak)
Philadelphia (Philly)

Figure 2: Alternative names used in lyrics.

In addition, resolving the ambiguity of toponyms (known as toponym resolution) in the lyrics is challenging. There have been many studies on toponym resolution [1, 11, 12, 21, 23, 35]. However, lyric data mostly come from a few main websites, and the sources usually cover the most popular songs around the world. Unlike song lyrics, news stories are presented and written in different ways from a journalist's standpoint. Each news source has its preferred words and topics, and thus a local lexicon of a news source can be built to help determine where the news is most likely from. In contrast, song lyrics from different sources are identical except with minute differences in annotation styles. As the word usage each lyric source uses is very similar, building a local lexicon for each source cannot benefit the determination of the spatial scope of song lyrics.

Another research challenge is to infer the correct place from limited geographical and textual evidence in the lyrics. For example, "Penny Lane" from the Beatles song refers to Penny Lane, a street in Liverpool, but the lyrics only mention barber shops, firemen, and the Queen but leaves no clue about any other placenames. Thus the intricacies lie on how to determine which Penny Lane is the one to which the lyrics refer from over 3 hundred records of Penny Lane from the gazetteer. In addition, the Penny Lane to which the song of "Penny Lane" refers does not appear in the first 50 results, and thus every interpretation seems to fit in the context without any additional metadata. We address the issue by using auxiliary artist data to resolve the ambiguity of toponyms.

Finally, we provide a map query interface to allow users to explore mentions of interest on the map. The user can browse a Beatles song map based on one song or the entire collection and be informed about the most mentioned keywords around a specific place. We believe that the system will be of help to music fans to understand the musical characteristics of a place.

2 RELATED WORK

In this section, we review the techniques used for building the MusicStand server and briefly discuss the existing similar systems using map query interface. This description is aided by making frequent references to the NewsStand system [10, 12, 24, 25, 27, 31] for reading news with a map query interface.

2.1 Truecasing

Truecasing is a means of restoring the true cases of tokens when the capitalization of words makes a difference. Lita *et al.* [14] propose a truecaser based on language modeling with sentence-level decoding to find the best well-cased sentence. Similar work in recovering case information of automated speech transcripts include Batista *et al.* [3] and Gravano *et al.* [8] using n-gram language models, Susanto *et al.* [30] using character-level recurrent neural networks, and Nguyen *et al.* [19] with the popular Transformer [32] model in NLP.

2.2 Named Entity Recognition

Named entity recognition (NER) is a fundamental task for information extraction where each word is classified in predefined categories. The extracted information can be further processed to be used in various fields such as machine translation, complex question-answering systems, and spoken dialog systems.

Traditional NER methods may run into trouble when geotagging lyrics with out-of-vocabulary words. It is because the methods are usually trained on a well-edited corpus, such as news articles or Wikipedia articles, which are very different from lyrics in terms of writing styles and grammar. Researchers have addressed the issues by decomposing out-of-vocabulary (OOV) words into subwords and characters so that the model can generate character-based and subword embeddings for OOV words.

The recent striking achievements in natural language processing (NLP) research lead to significant improvements in NER tasks.

Deep-learning-based NLP models have become increasingly dominant because of the ability to learn intricate data by the nonlinearity of activation functions. Especially after [32] came out, many transformer-based NLP models [4, 16, 36] achieved the state-of-art. Despite new robust models have sprung up in NLP domains, prior work indicate [14, 17] the performance of current NER systems plummet on badly-cased or non-cased text that are trained on standardly-edited datasets. There is limited literature on addressing this issue. The common solution is to infer the true case information of text being processed by a truecaser prior to feeding into NER systems. Recent work includes Mayhew *et al.* [17] combining a pre-trained truecaser with a BiLSTM-CRF model.

2.3 Geotagging

Geotagging is the process of identifying the geographical locations of a given content or photograph, video, etc. It has been broadly applied and studied due to the increasing prevalence of geotagged data. Unfortunately, not all geotags are directly accessible and require additional processing to acquire latent geographic information. The most common approach [1, 20, 22, 33] consists of two steps: (1) Toponym recognition and (2) Toponym resolution.

2.3.1 Toponym Recognition. Toponym recognition finds all geo-related words in a given text. Most existing toponym recognition systems utilize NER and Part-Of-Speech (POS) tagging to identify geo entities from a given text. [6, 15] studied how to identify toponyms from abbreviated, misspelled, or localized words on social media that were normally missed by traditional geoparsers. Leidner *et al.* [9] surveyed methods commonly used for toponym recognition, e.g., the gazetteer lookup-based methods by looking up a geographical dictionary, rule-based methods like context-free grammars or matching regular expressions, machine learning-based methods by learning from a gold standard corpus to recognize real-world toponyms.

2.3.2 Toponym Resolution. Toponym resolution determines the final geographic coordinates of the places if any ambiguity exists. Many researchers have studied this issue [1, 12]. Early prominent work in 2004, Amitay *et al.* [1] proposed applying heuristics to disambiguate toponyms with a focus-finding algorithm to picks foci by score. However, the gazetteer only contains about 40,000 prominent places with a population over 5,000 where sparsely-populated places are certainly missing. Lieberman *et al.* [12] consider dateline toponyms, hierarchical containment relationships, news sources' local lexicons, and a global lexicon to resolve ambiguous toponyms. Baldridge *et al.* [35] proposed using a hierarchy of logistic regression classifiers to geotag text by mapping text to discrete grid cells over the Earth's surface.

2.4 Systems Using a Map Query Interface

In this subsection, we briefly discuss existing similar systems developed for finding interesting topics using a map query interface.

2.4.1 NewsStand.

NewsStand [12, 31] is a system that enables users to search for news geographically. The system polls thousands of news sources for updates, translates articles in foreign languages, geotags news articles, and clusters news by location and news keyword. In addition, users can specify the layer of or the topic of interest to filter news, e.g., disease layer, brand layer, sci-tech topic, etc.

2.4.2 TwitterStand.

TwitterStand [28] is a system that presents breaking news in a timely fashion with a map query interface. The system first handpicks tweeters that are known to publish news, maintains a set of the most active news tweeters, and scrapes news tweets from the handpicked tweeters, the active tweeters, and a sampling of all the tweets. Then, the system filters out noisy tweets not related to news, geotags news tweets, and clusters news tweets that share similar stories.

2.4.3 Radio Garden.

RadioGarden [5] is an online radio station that allows users to tune in to a live radio station all over the world by rotating and zooming in and out of the globe. Not that the system is not fully considered as using a map query interface as it loads the full list of live radio stations all at once, which is not practical for a system with millions of records.

2.4.4 Listening Together.

Listening Together [29] provides a 3D interactive Earth globe to explore serendipitous encounters where the song is being played at nearly the same time by two different listeners. However, the system can't let you zoom in on a specific area of interest and automatically switches songs at random every few seconds.

3 MUSICSTAND ARCHITECTURE

The architecture of MusicStand consists of four main modules: (1) Scraper, (2) Truecasing, (3) NER, (4) Geotagging. Figure 3 presents an overview of the MusicStand pipeline. First, the lyrics are extracted from the raw HTML contents from the scraper and transformed into properly capitalized lyrics by a truecaser. Subsequently, the named entities in the lyrics are identified by NER, and then the geographic coordinates of the geographic named entities are determined by toponym resolution. Finally, the results are stored for various types of queries presented by a map query interface.

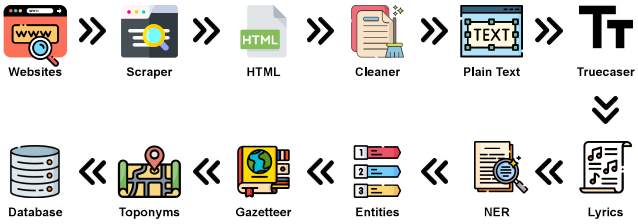

Figure 3: MusicStand architecture.

4 QUERIES

We have three types of queries, based on the nature of the search key, corresponding to "who", "what", or "where".

(1) who: display a marker on the map at each location mentioned in the lyrics of all songs performed by the search keyword.
(2) what: display a marker on the map at each location mentioned in the song with the search keyword title.

(3) where: display a list of titles of all songs whose lyrics mention the search keyword location.

In query 3, a single click on a result title R results in placing a marker on the map at each location mentioned in the lyrics of the song (equivalent to query 2 "what"), while a double click on a result title results in invoking YouTube to play the song.

Notice that all queries return a location on a map in one step with the exception of the "where" query 3 which requires two steps. Thus the main point of MusicStand can be stated as adding "who" to the "what" and "where" of NewsStand [26], PhotoStand [24], TwitterStand [28], and STEWARD [13] In all queries, if the number of song titles or locations gets too large, then the items (including multiple performers of the same song are ranked in terms of the total YouTube views of the relevant song. Other notable features:

(1) World map icon key yielding a map of the world with markers at the most prominent locations in terms of total YouTube views of the songs associated with them.
(2) "Local" key icon yielding a map of the area local to the user with the most prominent locations in terms of total YouTube views of the songs associated with them.
(3) "Home" key icon with the same type of map as in 1–2.

5 CONCLUSION

To the best of our knowledge, MusicStand(`http://musicstand.umiacs.io/`) is the first work to geotag lyrics and to present songs by geography. It has been an unsolved research topic for decades and still leaves room for improvement. Future plans include improving toponym recognition on microtext and combining machine learning-based methods for toponym resolution in the absence of evidence. We also plan to build a spatiotemporally-varying MusicStand system, which we believe is of critical use to explore the evolution of musical composition around a certain geographical area. We want to recognize historical geographical mentions. For example, currently "Penny Lane" is facing a name change, and there is a possibility that the mention cannot be identified in the future. It would be of interest to trace songs from hundreds of years ago.

ACKNOWLEDGMENTS

We are grateful to Mahmoud Sayed for suggesting MusicStand as an extension of NewsStand. This work was sponsored in part by the NSF under Grants IIS-18-16889, IIS-20-41415, and IIS-21-14451.

REFERENCES

[1] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. 2004. Web-a-where: Geotagging web content. In *SIGIR '04*. 273–280.
[2] AZLyrics. 2021. *AZLyrics*. `https://www.azlyrics.com/`
[3] F. Batista, D. Caseiro, N. Mamede, and I. Trancoso. 2008. Recovering capitalization and punctuation marks for automatic speech recognition: Case study for Portuguese broadcast news. *Speech Commun.* 50, 10 (10 2008), 847–862.
[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171–4186.
[5] Netherlands Institute for Sound and Vision. 2021. *Radio Garden*. `http://radio.garden/`
[6] J. Gelernter and S. Balaji. 2013. An algorithm for local geoparsing of microtext. *GeoInformatica* 17, 4 (Oct. 2013), 635–667.
[7] Genius. 2021. *Genius*. `https://genius.com/`
[8] A. Gravano, M. Jansche, and M. Bacchiani. 2009. Restoring punctuation and capitalization in transcribed speech. In *ICASSP '09*. 4741–4744.
[9] J. L. Leidner and M. D. Lieberman. 2011. Detecting geographical references in the form of place names and associated spatial natural language. *ACM SIGSPATIAL Special* 3, 2 (July 2011), 5–11.
[10] M. D. Lieberman and H. Samet. 2012. Supporting rapid processing and interactive map-based exploration of streaming news. In *GIS '12*. 179–188.
[11] M. D. Lieberman, H. Samet, and J. Sankaranarayanan. 2010. Geotagging: Using proximity, sibling, and prominence clues to understand comma groups. In *GIR '10*. Article 6.
[12] M. D. Lieberman, H. Samet, and J. Sankaranarayanan. 2010. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *ICDE '10*. 201–212.
[13] M. D. Lieberman, H. Samet, J. Sankaranarayanan, and J. Sperling. 2007. STEWARD: Architecture of a spatio-textual search engine. In *GIS' 07*. Article 25.
[14] L. Lita, A. Ittycheriah, S. Roukos, and N. Kambhatla. 2003. TRuEcasIng. In *ACL '03*. 152–159.
[15] F. Liu, M. Vasardani, and T. Baldwin. 2014. Automatic identification of locative expressions from social media text: A comparative analysis. In *Proceedings of the 4th International Workshop on Location and the Web (LocWeb '14)*. 9–16.
[16] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR* abs/1907.11692 (2019). `http://arxiv.org/abs/1907.11692`
[17] S. Mayhew, G. Nitish, and D. Roth. 2020. Robust named entity recognition with truecasing pretraining. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, Feb, 2020*, Vol. 34. 8480–8487.
[18] Musixmatch. 2021. *Musixmatch*. `https://www.musixmatch.com/`
[19] B. Nguyen, V. B. H. Nguyen, H. Nguyen, P. N. Phuong, T.-L. Nguyen, Q. T. Do, and L. C. Mai. 2019. Fast and accurate capitalization and punctuation for automatic speech recognition using transformer and chunk merging. In *2019 22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*. 1–5.
[20] B. Pouliquen, M. Kimler, R. Steinberger, C. Ignat, T. Oellinger, K. Blackler, F. Fluart, W. Zaghouani, A. Widiger, A.-C. Forslund, and C. Best. 2006. Geocoding multilingual texts: Recognition, disambiguation and visualisation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. 53–58.
[21] G. Quercini, H. Samet, J. Sankaranarayanan, and M. D. Lieberman. 2010. Determining the spatial reader scopes of news sources using local lexicons. In *GIS '10*. 43–52.
[22] M. A. Radke, N. Gautam, A. Tambi, U. A. Deshpande, and Z. Syed. 2018. Geotagging text data on the web—A geometrical approach. *IEEE Access* 6 (2018), 30086–30099.
[23] H. Samet. 2014. Using minimaps to enable toponym resolution with an effective 100% rate of recall. In *GIR '14*. Article 9.
[24] H. Samet, M. D. Adelfio, B. C. Fruin, M. D. Lieberman, and J. Sankaranarayanan. 2013. PhotoStand: A map query interface for a database of news photos. *PVLDB* 6 (08 2013), 1350–1353.
[25] H. Samet, M. D. Adelfio, B. C. Fruin, M. D. Lieberman, and B. E. Teitler. 2011. Porting a web-based mapping application to a smartphone app. In *GIS '11*. 525–528.
[26] H. Samet, J. Sankaranarayanan, M. D. Lieberman, M. D. Adelfio, B. C. Fruin, J. M. Lotkowski, D. Panozzo, J. Sperling, and B. E. Teitler. 2014. Reading news with maps by exploiting spatial synonyms. *Commun. ACM* 57, 10 (Oct. 2014), 64–77.
[27] H. Samet, B. E. Teitler, M. Adelfio, and M. D. Lieberman. 2011. Adapting a map query interface for a gesturing touch screen interface. In *WWW '11*. 257–260.
[28] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. 2009. TwitterStand: News in tweets. In *GIS '09*. 42–51.
[29] Spotify. 2021. *Listening Together*. `https://listeningtogether.atspotify.com/`
[30] R. H. Susanto, H. L. Chieu, and W. Lu. 2016. Learning to capitalize with character-level recurrent neural networks: An empirical study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2090–2095.
[31] B. E. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling. 2008. NewsStand: A new view on news. In *GIS '08*. Article 18.
[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *NIPS'17*. 6000–6010.
[33] R. Volz, J. Kleb, and W. Mueller. 2007. Towards ontology-based disambiguation of geographical identifiers. In *WWW '07*, Vol. 249.
[34] M. Wick and B. Vatant. 2021. *The geonames geographical database*. `https://www.geonames.org/`
[35] B. Wing and J. Baldridge. 2014. Hierarchical discriminative classification for text-based geolocation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 336–348.
[36] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. `http://arxiv.org/abs/1906.08237`