The Language of Jazz: A Natural Language Processing-based Analysis of the Patterns and Vocabulary of Jazz Solo Improvisation

Saebyul Park(a), Juhan Nam (b)

(a) KAIST, Korea, saebyul_park@kaist.ac.kr, (b) KAIST, Korea, juhan.nam@kaist.ac.kr

This paper utilizes natural language processing (NLP) techniques to investigate the language of jazz and enhance our understanding of jazz solo improvisation. The study focuses on applying the Byte-Pair Encoding (BPE) algorithm to tokenize jazz melodies and analyze the patterns and language associated with different aspects of jazz, including artist, style, and tempo. Using the Weimer Jazz Dataset, a jazz dictionary is created, and solo melodies are tokenized using sub-word dictionaries for comprehensive text analysis. Through empirical experiments and classification tasks, the effectiveness of tokenization is demonstrated by improving performance in various similarity algorithms and evaluation methods. Additionally, text analysis visualizations provide valuable insights into the distinct language and artistic expression present in different jazz solo melodies. By leveraging NLP algorithms and text analysis techniques, this study contributes to a deeper understanding of the musical patterns within the jazz language as a textual vocabulary, enabling further exploration of higher-level cognitive processes involved in musical creativity and artistic expression.

Keywords: jazz language, Natural Language Processing (NLP), Byte-Pair Encoding (BPE)

1. Introduction

As Louis Armstrong remarked, jazz can be understood as a form of language (Armstrong, 1964). This notion highlights the shared characteristics between jazz and language from a linguistic perspective. Indeed, music and language exhibit significant commonalities (Patel, 2003). Both possess hierarchical structures and utilize sequences for communication (Lerdahl & Jackendoff, 1996; Patel, 2003). They are intricate systems that enable the combination of smaller units, such as notes in jazz and morphemes in language, to generate a wide range of more intricate structures (Chomsky, 1965). Consequently, researchers have undertaken various studies to gain linguistic insights into musical patterns and structures (Lerdahl & Jackendoff, 1996; Narmour, 1990; Patel, 2003).

Jazz improvisation, in particular, often reveals discernible shapes and patterns that contribute to coherence and establish connections through a shared vocabulary (Norgaard, 2012). This observation has sparked numerous studies exploring the linguistic aspects and patterns within jazz solos, employing diverse methodologies encompassing neuroscience, linguistics, and computational analysis (Beaty, 2015; Donnay, Rankin, Lopez-Gonzalez, Jiradejvong, & Limb, 2014; Limb & Braun, 2008). However, despite the significant advancements in NLP that have expanded our understanding of language, the application of NLP techniques to analyze jazz as a form of textual or symbolic language has been relatively limited.

To address this gap, our study takes a data-driven approach to analyze the language of jazz in its textual format using NLP methods. We begin by defining a jazz dictionary that captures the unique vocabulary of jazz within the framework of NLP by employing the widely-used tokenization method of NLP, specifically Byte-Pair Encoding (BPE) (Sennrich, 2015). The Weimer Jazz Dataset (Pfleiderer et al., 2017), a collection of jazz solo transcriptions, is utilized to create the jazz dictionary, which is subsequently employed to tokenize jazz solo melodies for empirical analysis using text analysis techniques. Additionally, a classification evaluation is conducted to explore the characteristics of solo improvisation across different classes, including performance, style, and tempo. We employ different similarity algorithms to examine the effectiveness of classification in capturing the distinct features of jazz solo improvisation. Furthermore, by employing a variety of text-analysis tools and visualizations, we present a comprehensive and in-depth exploration of our research findings, enriching the understanding and interpretation of the results. By utilizing NLP techniques, this study contributes to the investigation of musical patterns in jazz language, shedding light on the distinct characteristics and qualities of jazz solo performances across different aspects, such as artist, style, and tempo.

2. Methods

2.1.Dataset

We utilize the Weimar Jazz Dataset (Pfleiderer et al., 2017) in this study to examine and analyze jazz improvisation. This dataset is a comprehensive collection of jazz solo transcriptions that encompasses a wide range of jazz styles and features performances by various artists. It provides researchers with a rich and diverse resource for studying and analyzing jazz improvisation, allowing for investigations into the linguistic aspects, melodic structure, rhythmic patterns, and harmonic elements of jazz solos. By leveraging this dataset, we identify frequently occurring words (or patterns) in jazz solos and conduct a classification experiment using the provided metadata.

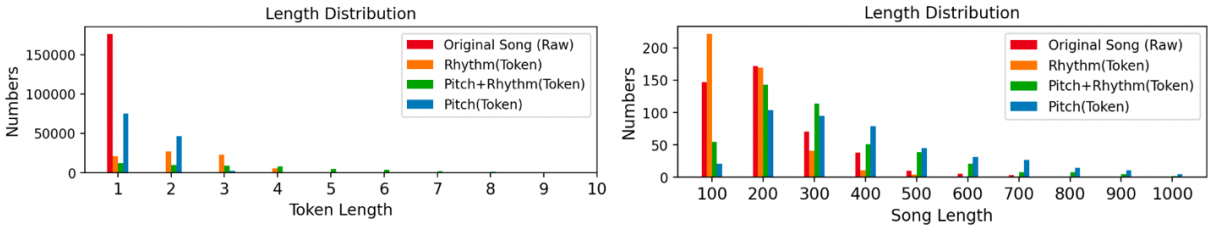

Figure 1. Statistical Distribution and Visualization of Melodies in the Weimer Jazz Dataset: Token Length and Song Length by Feature

2.2.Textual Representation

To analyze melodies in this study, we considered two key features: pitch and rhythm. The pitch feature was encoded using MIDI numbers, while the rhythm feature is extracted with inter-onset intervals (IOI) quantized by sixteenth notes. To represent these features as text, we converted them into 3-digit numbers and transformed them into strings. For the pitch feature, we ensured consistent formatting by adding a leading zero for numbers below 100 (e.g., MIDI number 60 representing middle C was represented as 060). Similarly, the IOI values were multiplied by 100 and converted into three-digit numbers (e.g., 1 beat was represented as 100). These individual features could be analyzed separately or combined into a single unit for text analysis purposes (e.g., the combined representation of pitch and IOI for a note would be represented as 060100).

2.3.Dictionary Generation

In this study, we utilized BPE to create a sub-word dictionary and analyze the language of jazz melodies. BPE is a robust technique that breaks down sentences and words into smaller units, capturing both morphological and semantic information. It starts with an initial vocabulary of individual characters and gradually merges the most frequent pairs of units to construct a comprehensive vocabulary that encompasses the linguistic features of the text. BPE has been widely adopted in various NLP tasks to improve performance and effectively handle out-of-vocabulary words. For our analysis, we treated each jazz melody as a sentence, with each note representing a character. By applying BPE to the Weimer Jazz Dataset, we iteratively merged the most frequent notes and phrases until no more repeating pairs were found. During the merging process, the '_' symbol was used to indicate the combination of two notes (e.g., '067_068' when notes '067' and '068' were merged). This process resulted in the creation of 4585 sub-words for the pitch feature, 2688 sub-words for the rhythm feature, and 5915 sub-words for the combined representation of pitch and rhythm features. This sub-word dictionary will be utilized for tokenizing the jazz melodies in subsequent analyses.

2.4.Melody Tokenization

Tokenization in this study refers to the process of dividing jazz melodies into smaller units, referred to as *tokens*, for the purpose of text analysis. The sub-word dictionary created earlier was utilized to tokenize the melodies, with a preference given to longer sub-words and those with higher frequencies. To ensure the inclusion of informative linguistic units, only sub-words that appeared in the dataset at least 10 times were considered for tokenization. Specifically, for pitch, we used 1213 sub-words; for rhythm, we used 727 sub-words; and for the combined feature, we used 1354 sub-words. This tokenization process enables more granular analysis of the jazz language, allowing us to explore the patterns, vocabulary, and linguistic characteristics of jazz solo improvisation in a more detailed manner.

3. Preliminary Data Analysis

3.1. Statistic Distribution

Figure 1 shows the statistical distribution of melodies obtained through the tokenization process of the Weimer Jazz Dataset. The left panel illustrates the token lengths for each feature, revealing that raw songs consist of single characters, while tokenized songs exhibit varying lengths. Specifically, the rhythm feature contains the longest words, followed by pitch and the combined feature (pitch and rhythm). The right panel provides a visualization of the resulting song lengths per feature. Remarkably, tokenization simplifies and shortens the melodies compared to the original raw songs. This demonstrates the effectiveness of the proposed method in transforming and simplifying melodies, facilitating a data-driven approach to meaningful text analysis.

3.2.Illustrative Examples

Figure 2 displays illustrative examples of the most frequent tokens found in the Weimar Jazz dataset. The top panel presents the top 15 tokens for each feature, excluding single notes. In the bottom panel, the five most frequently occurring tokens are shown for the combined feature, along with their corresponding score notes. The tokens with a length of 4 are displayed on the left side, while the tokens with a length of 3 are presented on the right side.

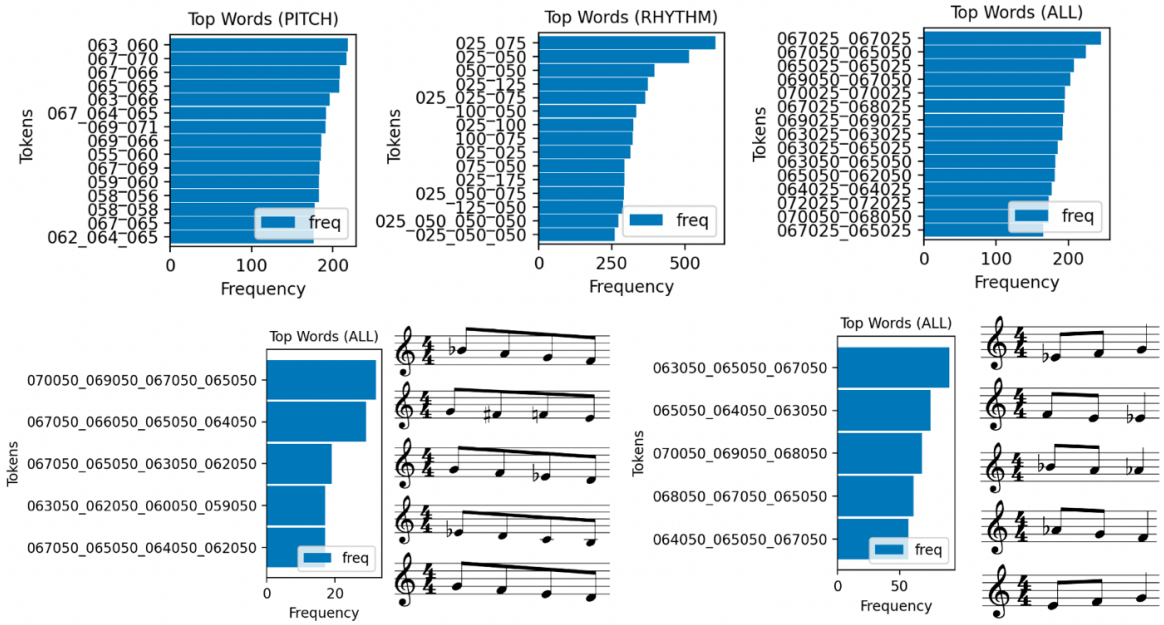

Figure 2. Illustrative Examples of Most Frequent Tokens in the Weimar Jazz Dataset

It is interesting to note that jazz solo melodies frequently feature an abundance of phrases that incorporate chromaticism, sharps, and notes beyond the confines of the traditional tonal scale centered around C chords. This prevalence of non-diatonic tones and extended harmonies in jazz improvisation enhances the melodic language of jazz, introducing complexity and richness. Additionally, an intriguing observation is that the top five longest phrases in the melodies all follow a top-down direction. This directional pattern suggests a characteristic melodic contour commonly employed in jazz improvisation.

4. Experiments

4.1. Task: Melody Classification

The empirical evaluation in this study focused on a melodic classification task aimed at observing and assessing the characteristics of jazz solo improvisation. The task involves the identification and classification of various aspects such as artist, style, and tempo using the metadata of the dataset. To ensure an adequate sample size, only artists, styles, and tempos with more than 10 songs were included in the classification task. Table 1 provides an overview of the classes used in the classification task.

4.2.Text Similarity Algorithms

In this experiment, three different similarity algorithms were employed to analyze and compare textualized melodic data. These algorithms include TF-IDF, Tversky Similarity, and Word2Vec.

TF-IDF similarity is based on the Bag-of-Words hypothesis and calculates document similarity using the TF-IDF representation of words. It takes into account the frequency of words in a document and their rarity across the corpus, allowing for meaningful comparisons.

Table 1. Overview of classification tasks and classes used in the study

Tasks	Classes	N
Artist	Charlie Parker, David Liebman, John Coltrane, Michael Brecker, Miles Davis, Sonny Rollins, Steve Coleman, Wayne Shorter	110 songs
Style	Bebop, Cool, Swing, Traditional, Postbop, Fusion, Hardbop	448 songs
Tempo	Up, Medium Up, Medium, Medium Slow, Slow	453 songs

Tversky similarity measure is a feature-based comparison approach that considers both common and distinctive features. It enables capturing specific characteristics and patterns in the melodic data, providing a flexible similarity measure.

Word2Vec vector similarity is based on the Distributional Hypothesis and represents words as dense vectors in a high-dimensional space. It captures the semantic meaning and relationships between words, allowing for the exploration of semantic similarities within melodies.

To implement the TF-IDF and Tversky measures, we utilized the Pystringmatching library (https://pypi.org/project/py-stringmatching/). In the case of Word2Vec, we utilized the Gensim library (https://radimrehurek.com/gensim/) to generate vector representations of the tokens. We calculated similarity between melodies by computing the average vector of the tokens and applying the cosine similarity measure. Table 2 provides a summary of the

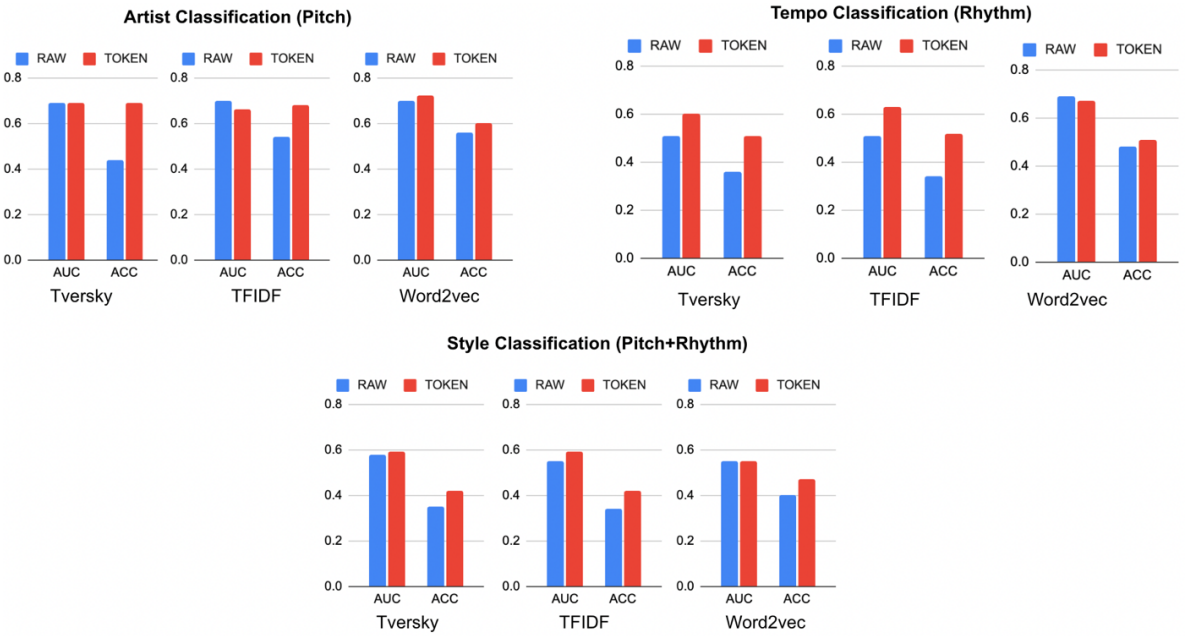

Figure 3. Classification Performance for Different Tasks Based on the Feature with the Highest Score

similarity algorithms used in this study. Further details and implementation about the methods can be found in the respective libraries.

Table 2. Summary of similarity algorithms used in the study

	Description
TF-IDF Similarity	Measures similarity between documents based on TF-IDF representation, capturing shared and rare terms.
Tversky Similarity Measure	Considers common and distinctive features between objects, allowing flexible comparison based on user-defined weights.
Word2Vec Vector Similarity	Calculates similarity between word vectors generated by Word2Vec, capturing distributional semantics.

4.3.Evaluation Metrics

The evaluation of the classification models in this study involved the use of two metrics: AUC (Area Under the Curve) and Accuracy (ACC). These metrics were employed to assess the performance and quality of the predicted results. AUC measures the ability of the models to discriminate between different classes, with a higher value indicating better discrimination. Accuracy, on the other hand, quantifies the proportion of correctly classified instances, providing an overall measure of the model's accuracy.

5. Results

5.1.Overall Results

Figure 3 provides an overview of the classification experiment results, highlighting the feature with the highest score for each task (pitch for artist classification, rhythm for tempo classification, and the combined pitch and rhythm for style classification). Melody tokenization consistently outperformed raw melodies on the majority of the task and evaluation measures, indicating the effectiveness of this approach in improving classification performance. The importance of different features varied across classification tasks, with the pitch feature being crucial for artist classification, the combined pitch and rhythm feature for style classification, and the rhythm feature for tempo classification. These findings emphasize the significance of each feature in capturing the unique patterns and characteristics of jazz improvisation.

In the artist classification task, Word2Vec achieved the highest accuracy (0.72), and the use of tokenization notably improved the performance of the Tversky measure by 25% (0.43 to 0.69), demonstrating the efficacy of tokenization in accurately identifying and classifying jazz artists.

For tempo classification, the proposed method generally improved the classification performance. However, it is worth noting that the highest score was achieved by the raw melody's AUC measure (0.63) using word2vec. This result can be attributed to the

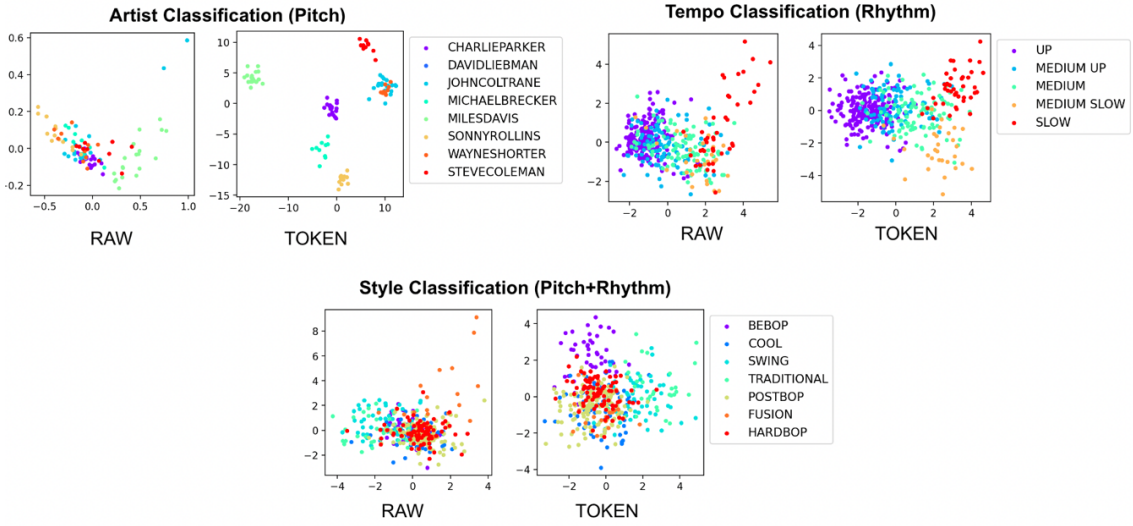

Figure 4. LDA Visualization of Artist, Style, and Tempo Classification

Figure 5. WordCloud Visualization of Artist-Specific Tokens with TF-IDF Weighting (Pitch Feature)

longer average length of the rhythm feature when tokenized compared to other features. Furthermore, since word2vec captures contextual information, it is suggested that the rhythm characteristics are well represented even in the raw melody without tokenization, as indicated by the AUC score that reflects the recall rate.

For style classification, although the performance may not have reached high levels (with a peak of 0.59 for the token result using the Tversky measure), it is significant to observe that tokenization consistently outperformed raw melody across all measures and evaluations. These results indicate that there is still room for improvement in this task, and further enhancements can be achieved by considering larger datasets and employing appropriate algorithms to enhance the classification performance.

5.2. LDA Visualization

Figure 4 presents the Latent Dirichlet Allocation (LDA) visualization results for the three classification tasks: artist classification, style classification, and tempo classification. The LDA visualization is a commonly used probabilistic topic modeling algorithm in text analysis, which enables us to visually explore and analyze the distribution of classes within the dataset. In this study, the LDA visualization was implemented using the average vector melodies obtained from the Word2Vec classification experiment.

In the artist classification task, where tokenization significantly improved the performance, the LDA visualization of tokenized melodies clearly shows distinguishable differences between different artists. This visual representation indicates the effectiveness of tokenization in capturing artist-specific patterns and characteristics.

For tempo classification, there is a noticeable distinction between classes in both raw and tokenized melodies, indicating that tempo information can be visually observed to some extent regardless of the tokenization process.

In the style classification task, where the overall performance is lower, distinguishing between classes remains challenging. However, the LDA visualization of tokenized melodies still provides visual cues that can aid in identifying patterns and characteristics associated with different styles.

5.3.WordCloud Visualization

Figure 5 presents a wordcloud visualization with TF-IDF weights applied to highlight the commonly used vocabulary by individual artists in the Weimer Jazz Dataset. A WorCloud is a popular visualization technique in NLP that visually represents the frequency of words in a corpus. In this case, the TF-IDF weight emphasizes the uniquely used phrases by each artist, providing insights into their distinctive vocabulary and linguistic style. The WordCloud visualization clearly illustrates the diversity and creativity in jazz solo improvisation, with each artist presenting their own language and artistic expression. This further emphasizes the importance of artist-specific analysis and understanding of jazz language.

6. Discussion

In this study, we utilized NLP techniques to investigate the language of jazz and enhance our understanding of jazz solo improvisation. Through the application of the BPE algorithm, we successfully tokenized jazz melodies, allowing us to analyze and identify patterns and language associated with different aspects of jazz, such as artist, style, and tempo. Furthermore, we conducted empirical experiments to examine jazz melodies based on these aspects, providing a comprehensive understanding of the unique language and characteristics of jazz improvisation. A notable finding of this study is the effectiveness of tokenization in improving classification performance. Across multiple tasks and evaluation metrics, tokenized melodies consistently outperformed untokenized melodies, showing the worth of this approach for capturing meaningful features. In addition, the use of multiple text analysis visualizations demonstrated the successful use of NLP techniques for visually analyzing class distributions within the dataset. These visualizations notably showed distinct differences between artists, thereby providing a valuable understanding of their individual languages and artistic expressions. Through the application of NLP algorithms and text analysis techniques, this study contributes to a better understanding of musical patterns within the jazz language as a textual vocabulary. We believe this research provides the groundwork for a deeper exploration of higher-level cognitive processes involved in musical creativity and artistic expression.

References

Armstrong, L. (1964). Jazz is a language. *Music Journal*, 22(1), 16.

Beaty, R. E. (2015). The neuroscience of musical improvisation. *Neuroscience & Biobehavioral*

Reviews, 51, 108–117.

Chomsky N (1965) *Aspects of the theory of syntax.* Cambridge, Massachusetts: The MIT press.

Donnay, G. F., Rankin, S. K., Lopez-Gonzalez, M., Jiradejvong, P., & Limb, C. J. (2014). Neural substrates of interactive musical improvisation: an fmri study of 'trading fours' in jazz. *PLoS one*, 9(2), e88665.

Lerdahl, F., & Jackendoff, R. S. (1996). *A generative theory of tonal music*. MIT press.

Limb, C. J., & Braun, A. R. (2008). Neural substrates of spontaneous musical performance: An fmri study of jazz improvisation. *PLoS one*, 3(2), e1679.

Narmour, E. (1990). *The analysis and cognition of basic melodic structures: The implication-realization model.* University of Chicago Press.

Patel, A. D. (2003). Language, music, syntax and the brain. *Nature neuroscience*, 6(7).

Pfleiderer, M., Frieler, K., Abeßer, J., Zaddach, W.-G., & Burkhart, B. (Eds.). (2017). *Inside the Jazzomat - New Perspectives for Jazz Research*. Schott Campus.

Sennrich, R., Haddow, B., & Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv*:1508.07909.