Conditional Text Generation with Force Word

Barbara Noémi Szabó
b.n.szabo@student.utwente.nl

Tobias Hoppe
t.hoppe-1@student.utwente.nl

Abstract

The goal of the project was to develop a model for a conditional sentence generator that only takes a word, given by the user together with the respective part-of-speech (POS) tag. As an exception the sentence generator should also generate a sentence in case to part-of-speech tag is not given at all. If this happens, the model assigns a probability to the most likely POS-tag. The structure, in which the word will be implemented is a simple subject-verb-object (SVO) structure. The project dives into the challenges faced during the implementation of the generator and how they were tackled as well as issues that still need to be solved in further iterations. In order to evaluate the generated sentences by the model, a language tool from Java as well as human feedback comes into place. The work presented in this report offers insights into the development and future challenges for text generators.

1 Introduction

In a world increasingly dominated by natural language processing, the ability to generate coherent and contextually relevant sentences is a crucial task. Our research project is centered on the development of a conditional sentence generator, a tool designed to generate sentences based on a specific word and its associated speech tag. The fundamental premise of this project is to create sentences that incorporate the provided word in a grammatically correct context, adhering to a simple subject-verb-object structure, enriched with optional adverbs for frequency, place, or time.

While the objective may seem straightforward, the challenges are multifaceted. We are tasked with constructing a sentence generator that not only accurately identifies the grammatical part of speech for a given word but also ensures that the resulting sentences make logical sense within the context of the chosen word. To this end, our project delves into several key questions:

1. *Grammatical Identification:* How can we reliably identify the grammatical part of speech and part of sentence for a given word? This is vital for creating sentences that conform to the rules of syntax and semantics.

2. *Semantic Coherence:* How can we guarantee that the generated sentences are contextually coherent and meaningful? In essence, how can we ensure that the generated sentences "make sense" in a real-world context?

3. *Embeddings and Cosine Similarity:* Can we use word embeddings and cosine similarity to address words not found in our vocabulary? This approach is essential for handling a wide range of words and ensuring that the generator can adapt to various linguistic inputs.

4. *Forced Inclusion of a Given Word:* How can we develop a system that compels the model to generate sentences that contain a specific word without violating the principles of grammar and meaning?

To address these questions, our project has employed a probabilistic model that is not AI-based, thus focusing on a different approach than contemporary AI-driven language models. We have also drawn on diverse corpora, including running text and treebanks, for training our model, striving to create a versatile and effective sentence generator.

This project represents an attempt to bridge the gap between linguistic understanding and computational sentence generation. By exploring these questions and developing a practical solution, we aim to contribute to the field of natural language processing and facilitate the generation of coherent and contextually relevant sentences for a wide range of applications.

Our model's capabilities can be readily explored and experienced through a user-friendly interface, it can be used after installing the code and the requirements from github https://github.com/TheBarbaraIsTaken/SentenceGenerator. This interactive platform allows users to input specific a word, choose its desired speech tag, and witness our conditional sentence generator in action. It provides a convenient means to experiment with sentence generation, experiment the model's proficiency in adhering to grammatical rules, and observe if the sentences produced are contextually meaningful. We invite you to visit the website and engage with our linguistic tool, fostering a deeper understanding of its capabilities.

2 Related Work

In previous research, grammatical models have already been used to provide unsupervised parsing based on the linguistic notion of a constituency test. Therefore, a set of sentences will be transformed by adding words intentionally. Afterwards, these sentences shall be judged towards their grammatical correctness (Cao et al., 2020). Sentence generation has been done quite frequently in the past with the help of Markov chains and hidden Markov models (Almutiri and Nadeem, 2022). Due to time constraints we have decided to perform the sentence generation based on the bigram probabilities gained from a corpus. This approach also has been of proof in recent research (Oh and Rudnicky, 2002). Another approach used bigrams, especially the bigram semantic distance, as index for the continuous semantic flow of generated text. This approach shows the relevance of birgram probabilities in times of neural networks (Reilly et al., 2023). Overall, bigrams seem to be a good option when sequences not necessarily of the same length (Pandey and Roy, 2023). Therefore, for our project approach we decided to work with probabilities based on co-occurrence.

3 Data

We incorporated the Universal Dependencies English Web Treebank (UD_English-EWT) version 2.12 (Bies et al., 2012) as a fundamental element of our dataset. This corpus represents a gold standard for universal dependencies in the English language and was derived from the source material found in the English Web Treebank LDC2012T13. While developing our dataset, we also explored the use of plain text and annotated it with spaCy Linguistic Features. Ultimately, we opted to utilize the UD_English-EWT corpus to ensure the quality and accuracy of our training data, prioritizing a low error rate in linguistic features.

The dataset comprises a substantial collection of 254,825 words, organized into 16,621 sentences. These sentences are drawn from various genres of web media, including weblogs, newsgroups, emails, reviews, and Yahoo! Answers, offering a diverse linguistic landscape. Detailed information on the sources of these sentences is available in the documentation.

It underwent a meticulous annotation process. Initially, the dependency trees were automatically converted into Stanford Dependencies and subsequently hand-corrected to Universal Dependencies. All basic dependency annotations have been single-annotated, with a limited portion double-annotated to enhance annotation consistency. Additional aspects of the treebank, such as Universal POS (part of speech), features, and enhanced dependencies, were primarily generated automatically, with minimal manual correction.

The Universal Dependencies English Web Treebank annotations are copyrighted material, extending from 2013 to 2021, attributed to The Board of Trustees of The Leland Stanford Junior University. These annotations are licensed under a Creative Commons Attribution-ShareAlike 4.0 International License, allowing for open and collaborative use.

It is structured as a collection of sentences annotated using Universal Dependencies annotation, and it contains a diverse range of linguistic expressions found in web media. The dataset conforms to the CoNLL-U format defined for Universal Dependencies and is compatible with the Universal Dependencies taxonomy.

4 Method

4.1 Predicting POS for The Given Word

The first step is the prediction of the part of speech (POS) for the given word if it is not provided by the user. In such cases, our model employs a default strategy wherein it assigns the most frequent POS tag associated with the given word. This approach ensures that even when specific POS information is absent, our sentence generator can make informed grammatical decisions based on the word's historical usage, enhancing the overall quality and coherence of the generated sentences. By addressing this inherent challenge, we aim to provide a robust and user-friendly solution for a wide range of inputs.

4.2 Predicting SVO Tag for The Given Word

In our methodology, predicting the Subject-Verb-Object (SVO) tag for the given word plays a pivotal role in constructing sentences. To achieve this, we employ a multifaceted approach that takes into account both explicit and implicit syntactic relationships in the Universal Dependencies English Web Treebank. Our goal is to assign the appropriate SVO tag to a word, ensuring that it fits seamlessly into the generated sentence structure.

We use a rule-based, basic assignment for verbs: if the user specifies that the given word is a verb, we assign the <VERB> tag directly. This step is straightforward, as it involves a direct one-to-one mapping between the user's input and the SVO tag.

For words that are not explicitly tagged as verbs by the user, we turn to the data within the corpus. We extract subjects and objects from the corpus along with their corresponding POS tags, encompassing not only words directly associated with subject or dependency labels - for example "nsubj" (nominal subject) or "dobj" (direct object) - but also the entire subtrees from the dependency parse. This approach allows us to use a probabilistic method also when we assign a tag for articles or adjectives.

We adopt this approach to be able to use relative frequencies of words associated with POS tags. By examining the entire syntactic context in which a word appears, we can capture the relationships between words and their roles in various sentence structures. This analysis ensures that the probabilistic assignment of SVO tags is grounded in a more holistic understanding of how words function within the language. As a result, our model can make informed decisions when determining SVO tags for words enhancing contextual relevance of the generated sentences. For example, if a word frequently occurs as a subject in similar sentence structures, it is more likely to be tagged as a <SUBJECT>.

By incorporating both explicit and implicit syntactic relationships and utilizing the wealth of linguistic features in the corpus, our model can intelligently assign SVO tags to words, thus ensuring that they fit seamlessly into the generated sentences, whether they represent subjects, verbs, or objects.

4.3 Expanding The Parts

Expanding the subject and object within a sentence is the next aspect of our methodology, allowing us to create linguistically rich content. To accomplish this expansion, we employ a systematic approach that involves utilizing bigrams to generate a sequence of part-of-speech (POS) tags before and after the given word. The "None" tag is used to indicate the beginning and end of the sequence. Subsequently, we employ a Hidden Markov Model (HMM) to predict words for the given POS sequence, both preceding and following the specified word.

By integrating these techniques, we can dynamically expand subjects and objects in the sentence, ensuring that they seamlessly blend with the context and align with the linguistic rules governing the English language.

In our quest to enhance the expansion of sentence components, we initially explored the use of bigrams as a means of capturing contextual information. However, we found that relying solely on bigrams yielded less grammatically accurate results.

4.4 Generating Words for The Remaining Part of The Sentence

In the process of generating sentences, the last step revolves around predicting the words for the remaining part of the sentence, specifically, the verbs, subjects, and objects. To accomplish this, we employed a method akin to n-grams, designed to create sentences that adhere to the rules of syntax and semantics. Our approach hinges on several key assumptions that govern the dependencies between these linguistic elements:

Subject-Verb Dependency: We assume that the verb in a sentence primarily depends on the subject. This is particularly significant for ensuring proper verb conjugation, especially in present tense when dealing with third person singular subjects.

Verb-Object Dependency: Likewise, we assume that the object is mainly dependent on the verb. This is crucial because not all verbs require a direct object, and our model must account for this variation.

Our method distinguish the following situations:

1. Subject Provided: If the user specifies the subject, we predict the verb first from the subject and then the object from the predicted verb. This ensures that the generated sentence maintains the correct grammatical structure.

2. Verb Provided: When the user provides the verb, we predict the subject from the verb and the object from the verb. This approach guarantees that the generated sentence maintains coherence with the supplied verb.

3. Object Provided: In the case of an object being given, we predict the verb from the object, taking care to select a verb that is likely to require

an object. Additionally, we predict the subject from the verb to ensure that the conjugation remains correct.

To predict words for these elements, we constructed co-occurrence matrices, specifically subject-verb and object-verb co-occurrence matrices. Notably, we do not predict the subject from the object, or vice versa, as such dependencies are typically less frequent in English sentence structures so we don't need to take it into account the subject-object co-occurrences.

In cases where a word, such as a subject, is not found in the vocabulary, we leverage word embeddings to identify the most similar subject that is present in the vocabulary. Subsequently, we predict a word for this analogous subject. This approach enhances the robustness of our model when dealing with unknown or out-of-vocabulary words, improving the overall quality of the generated sentences.

Moreover, for verb prediction, we do not merely extract the root verb, as this could lead to grammatically incorrect sentences (e.g., "he going" instead of "he is going"). Instead, we extract the entire verb phrase, which includes the children of the verb from the dependency graph marked with "AUX" or "PART" POS tags, as well as verbs accompanied by the "xcomp" dependency relation, thereby encompassing modal verbs.

In our pursuit of an effective method for predicting verbs based on subjects, we initially explored the use of word embeddings, which have proven to be powerful tools for capturing semantic relationships in language. However, our experimentation with nominal pronouns, particularly in the context of simple present tense conjugation, revealed that the performance of this approach was less robust than anticipated. When we tested 140 verbs, we found that the accuracy for assign a higher similarity score (utilizing cosine similarity) to the correct subject is only 55.7% (which is really close to random selection between the two classes: first number third person pronouns and the other pronouns). This outcome highlighted the inherent challenges in relying solely on embeddings, prompting us to seek a more reliable solution.

That's why we ultimately opted to develop a probabilistic model based on co-occurrence patterns. This model takes into account the statistical relationships between subjects and verbs, leveraging the power of linguistic context and syntactic dependencies to make informed predictions. By building a model grounded in co-occurrences, we aim to enhance the accuracy and contextual relevance of the subject-verb pairs in our generated sentences, addressing the limitations observed with the embedding-based approach.

4.5 Supplement with Adverbs

In our quest to enhance sentence variety and linguistic richness, we sought to implement a solution for supplementing sentences with adverbs. Our approach involved utilizing the subject of the sentence to dynamically assign the most probable adverb based on co-occurrences observed in the corpus. While this method aimed to introduce nuanced details and improve the overall quality of the generated sentences, our experiments revealed that the implementation introduced more linguistic inaccuracies compared to the previous existing model. Consequently, we made the strategic decision to exclude this feature from the final version of our sentence generation model. In the forthcoming Conclusion and Discussion section, we will delve into alternative strategies and considerations that guided our decision-making process, providing insights into the complexity of integrating adverbs seamlessly within the context of natural language generation.

5 Experiment and Results

A common way to analyse generated sentences is calculating scores (BLEU or ROUGE). A problem for us to use these was that we don't have a ground truth reference to compare the generated sentence to. Furthermore, recent research suggests that these methods might not be as useful as they used to be. BLEU and ROUGE scores do correlate negatively with human judgements especially in terms of fluency. Additionally the methods only output a single score while the feedback given from a human participant is more extensive and informative (Sai et al., 2022).

That's why we decided to evaluate the output sentences in two ways.

Firstly, the sentences will be checked on grammatical correctness. In order to do this, we worked with the java package called 'language-tool-python'. After installing java and installing the module with pip, its' use is very straightforward. After introducing the tool for the respective language, English, just one line of code can check each sentence for mistakes.

For the same 100 features we generated sentences 100 times and calculated the number of grammatical errors for each sentence. The total number of errors in the 10 000 sentences are: 4 603, so it means that the average number of errors for a sentence is about 0.46. The number of perfect sentences is 6100, which means that more than 60 % of the sentences doesn't contain grammatical error.

The average of errors were zero for the following features: (piranha, NOUN); (direction, NOUN); (price, NOUN). The maximum of average errors were created for the following configuration (investors, NOUN). Please note, that for one of the grammatical errors we encountered pertains to our treatment of negation for verbs. In some cases, you may observe "ca n't" written instead of "can't" in the generated sentences. This discrepancy arises from the way the corpus separates the verb stem and the negation element, and we have yet to fully resolve this issue.

Secondly, in our ongoing efforts to enhance the coherence and relevance of generated sentences, we incorporated human feedback as a valuable component of the evaluation process. Acknowledging the time constraints, we engaged a panel of 7 participants to evaluate a set of 100 randomly generated sentences, a common practice in the field of Natural Language Generation (NLG) (van der Lee et al., 2021).

Each participant was instructed to judge the sentences based on their sense making as well as syntactical appropriateness.

The participants provided an average score of 2.98 for the generated sentences, indicating a moderate perception of quality. About the distribution of scores, image 1 illustrates a distribution that is close to being uniform with peaks at 1 and 5.

Examining the distribution of average scores for the sentences, as shown in image 2, reveals a spread of opinions, with the majority clustered around the middle range. This dispersion suggests variability in the perceived quality of sentences. The cause of this phenomenon might be that we randomly selected words from the test corpus. This means that words that are not that valuable such as "t2i," "w.," or "dp," have been chosen which leads to sentences that make less sense.

The results show that 41% of the sentences received an average score greater than or equal to 3, underscoring the ongoing commitment to improving the overall quality of our sentence generation model. This iterative feedback loop is crucial for addressing nuances and further optimizing the model for meaningful and coherent outputs.

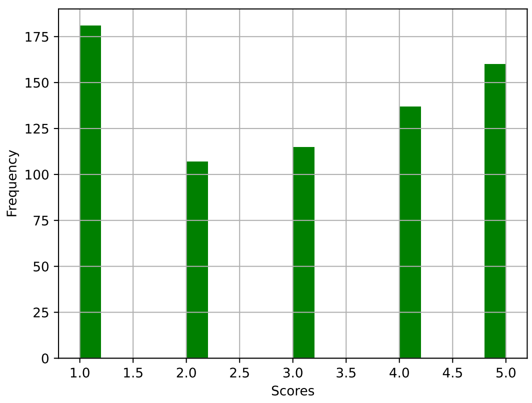

Figure 1: The distribution of ratings

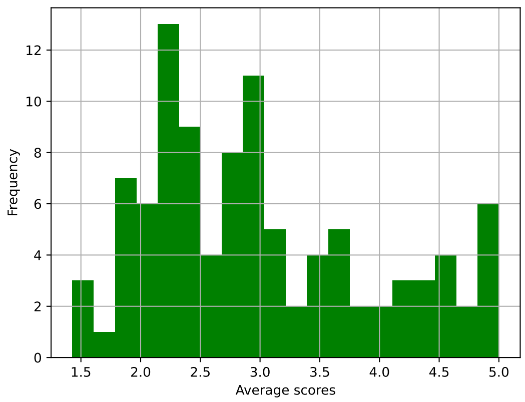

Figure 2: The distribution of average scores for sentences

In a lot of the generated sentences the adverb only has been added to the end of the sentence. This is not entirely how we planned it. We wanted to extend the sentence for the adverb but then the sentence should still go on further and not just stop. Sentences generated in this way were for example: Seth try to rally the great excess involved group significantly. He is doing some cognitive hidden ebook very. The first example is a correct one while the second should still be extended further in order to make sense.

6 Discussion

One of the issues of the program is that the predicted adverbs to extend the sentences mess up the grammatical as well as the syntactical correctness and accuracy of the program. The generated sentences start to make no sense anymore even though they did before extension of the sentences with adverbs. Therefore, this part had to be removed of the final version. There are different options to improve on this issue. For example, adding probabilities to the generation in order to let the function know which adverb might be meaningful at different sentence compositions might help. Another possibility is to extend the adverb even further, such that the sentence will not stop. An issue about this might be that the sentence will just stop to make any sense at all due to more words randomly added to it. Therefore, the conditional structure behind it has to be improved.

During the evaluation it became evident that it was quite hard for most of the participants to judge the given sentences purely on their sense making, disregarding issues within the grammar. Therefore, this is a factor to keep in mind when evaluating their judgement scores. In order to improve this, additional tests having participants judge grammar and sense making on two different scales might yield to more accurate results. Further on, the very extreme difference in the rating distribution could have been avoided with more participants. Also the insecurity on how to rate sentences might have yielded to either very positive ratings for only sentences, which are a hundred percent correct and very negative ratings regarding those with a lot of mistakes in them.

7 Conclusion

Overall, the project has successfully shown how a sentence generator can be implemented to generate a sentence, given only one word as input. Additionally to the word, the respective POS-tag can be given, but also in cases without any POS-tag the sentence generator is still able to generate good sentences in a lot of cases. The fact that our sentence generator yielded to more than 60% grammatically correct sentences is satisfying, considering that this is the first step into a problem like that.

Still there are many things to improve. The human feedback clearly shows, that even if more than half of the sentences are correct it does not necessarily mean that also of those sentences are also understandable to humans. This shows the limitations of rule based and probabilistic approaches.

References

Talal Almutiri and Farrukh Nadeem. 2022. Markov models applications in natural language processing: a survey. *Int. J. Inf. Technol. Comput. Sci*, 2:1–16.

Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. 2012. English web treebank. Web download, Linguistic Data Consortium, Philadelphia. Member Year(s): 2012.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Unsupervised parsing via constituency tests. *arXiv preprint arXiv:2010.03146*.

Alice H Oh and Alexander I Rudnicky. 2002. Stochastic natural language generation for spoken dialog systems. *Computer Speech & Language*, 16(3-4):387–407.

Abhishek Kumar Pandey and Sanjiban Sekhar Roy. 2023. Natural language generation using sequential models: A survey. *Neural Processing Letters*, pages 1–34.

Jamie Reilly, Ann Marie Finley, Celia P Litovsky, and Yoed N Kenett. 2023. Bigram semantic distance as an index of continuous semantic flow in natural language: Theory, tools, and applications. *Journal of Experimental Psychology: General*.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, 55(2):1–39.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.