Machine translation is ubiquitous in modern-day technologies. We interact with it on our phones, on television, and social media. Machine translation (MT) models are now predominantly trained using statistical techniques, which estimate the probability of translations using a combination large monolingual corpora and smaller parallel (sentence-aligned) corpora (bilingual or multilingual data).

Importantly for practical uses of MT, statistical biases that are encoded in machine translation models frequently generate harmful stereotypes. For example, consider the sentence, "The nurse sent the doctor the memo." In this sentence, you may have assumed that the nurse was a woman -- and this tendency is evident in demographics for that profession in the United States. However, [the gender identities of nurses without additional context are not guaranteed to be the gender stereotype of that profession](https://github.com/rudinger/winogender-schemas/blob/master/data/occupations-stats.tsv). Thus, it is important that machine translation systems preserve (grammatical) gender information where it exists, and ignore it where it does not exist.

This is particularly important for languages that do not mark gender information. For example, in languages like Turkish or Finnish, the pronouns that we may use to talk about he/she/they/xe/etc. are all translated as a single category (hän in Finnish; see table below).

In statistical natural language processing, this bias can be especially hard to overcome because modern systems have learned to associate professions with different pronouns on the basis of their co-occurrences. If we treat machine translation as a “mapping problem” we run into the following issue:

|  |  |  |
| --- | --- | --- |
| **English pronoun** | **Pronoun features** | **Finnish pronoun** |
| he/him/his/himself | third person singular, masculine | hän/hänen/häntä/../etc.  [Wikipedia page for "Finnish grammar"](https://en.wikipedia.org/wiki/Finnish_grammar)  [Wikipedia for Finnish "hän"](https://en.wiktionary.org/wiki/h%C3%A4n#Declension) |
| she/her/hers/herself | third person singular, feminine |
| they/them/their/theirs/themself/themselves | third person singular, unspecified/nonbinary/unknown |
| xe/xem/xyr/xemself | third person singular, neopronoun |

Many resources exist for studying gender bias in machine translation. For example, the Winogender dataset (<https://github.com/rudinger/winogender-schemas>) tests neural language model bias toward different gendered pronouns on the basis of the stereotyped profession of the noun in the sentence.

For this assignment we will use what is known as “round trip machine translation” to better understand grammatical gender in translation models. We will take sentences that have been used in parallel English-Finnish corpora that contain pronouns and we will assess the quality of the machine translated data on the following basis:

* Quality of machine translation – faithfulness between a “gold” translation and the translation model’s output using simple metrics
  + BLEU
  + ROUGE
* Faithfulness of pronoun usage between “gold” translation and model-translated output

*The main goals of this assignment are to:*

* Increase your familiarity with evaluation metrics for evaluating natural language generation and machine translation
* Increase your familiarity with some sources of errors in machine translation
* Better understand a social bias issue in natural language datasets

**Submission details**

We will ask you to load in the data and compute statistics. We will ask you to input numeric answers, upload your code for specific functions, and describe your solutions **on UBLearns** – note that **there is no autograder for this assignment**. Your answers will instead be associated with specific questions.

**Q1. Inspecting the subtitle round trip translation data (35 points)  
Goal: Evaluate the accuracy of translation**

1. Identify the column name for the data in eng-via-fin\_translations.tsv that contains the original "gold" text. Identify the column name that contains the text created by the round trip machine translation procedure.

Gold and Backtranslated

1. In the original text (the column containing the "gold" sentences), fill in the below table, noting that some sentences may contain any or all of these:

|  |  |
| --- | --- |
| Pronoun class | Proportion of sentences that contain this pronoun |
| he/him/his | 58.333333333333336% |
| she/her/hers | 18.333333333333332% |
| they/them/their/theirs | 10.833333333333334% |

Proportions of sentences containing pronouns:

he/him/his: 58.333333333333336%

she/her/hers: 18.333333333333332%

they/them/their/theirs: 10.833333333333334%

1. Qualitative evaluation: Describe 3 types of pronoun translation errors that you observe in the data and include an example of each of the errors.

Incorrect Pronoun Gender:

Original Text: "She acknowledged her mistake."

Machine Translation: "He confessed his mistake."

Explanation: In this example, the machine translation incorrectly changes the pronoun gender, replacing "she" with "he" without considering the original gender indicated in the source text.

Pronoun Omission:

Original Text - "I made her a doll."

Machine Translation - "I made a doll."

Explanation: Pronoun omission occurs when the machine translation fails to include the appropriate pronoun in the translated text. In this case, the pronoun "her" is omitted, resulting in an incomplete translation that lacks the indirect object of the sentence.

Incorrect Pronoun Agreement:

Original Text: "The room was light enough for her to read the letter."

Machine Translation: "The room was light enough for him to read the letter."

Explanation: This error occurs when the machine translation does not maintain proper agreement between the pronoun and its antecedent. In the example, the pronoun "her" should agree with the noun "room," but the machine translation incorrectly uses "him" instead.

1. Quantitative evaluation: Write a function that will compute scores for all of the sentences and store these scores in an array. Compute the mean scores for each.
   1. Unigram precision (related to BLEU)
   2. Unigram recall (related to ROUGE)

Unigram Precision: 0.7676816218911807

Unigram Recall: 0.7058646750438631

1. Neural evaluation: Using the sentence\_transformers package, compute
   1. the average (mean) cosine similarity between each of the Gold and Translated sentence pairs
   2. the number of translations that have a cosine similarity of 1

Average Cosine Similarity: 0.8849022

Number of Translations with Cosine Similarity of 1: 5

and answer the following questions:

* 1. When do we expect to see pairs of sentences with a cosine similarity of 1, in the context of sentence\_transformers?

In the context of sentence\_transformers, it is rare to see pairs of sentences with a cosine similarity of exactly 1. Cosine similarity ranges from -1 to 1, where a value of 1 indicates that the two vectors (representing the sentences) are identical in direction. However, due to the nature of vector representations and the inherent variability in natural language, it is uncommon for two different sentences to have an exact cosine similarity of 1.

While it's theoretically possible to have a cosine similarity of 1 between two sentences, it typically occurs in cases where the sentences are very similar or nearly identical, with minimal variation in their wording. In most practical scenarios, even highly similar sentences will have a cosine similarity slightly less than 1. Therefore, in the context of sentence\_transformers, it's not common to expect a large number of sentence pairs with a cosine similarity of 1.

* 1. Name one advantage and one disadvantage of using sentence\_transformers over ngram-based methods like BLEU or ROUGE

Sentence\_transformers can capture semantic similarity and contextual information, which gives them an edge over n-gram-based approaches like BLEU or ROUGE. Sentence\_transformers can better capture the meaning and context of sentences by encoding sentences into dense vectors using transformer models. This enables a more complex assessment of phrase similarity or translation quality that takes into account elements other than merely n-gram matching.

Sentence\_transformers have the drawback of requiring a lot of computational time and resources for training and inference. In comparison to n-gram-based techniques, transformer models are often more complicated and expensive to compute. When working with huge datasets or several languages, sentence\_transformers models may need to be trained on high-performance hardware and subjected to lengthier processing times.

Additionally, sentence\_transformers models rely on pretraining on large corpora, which means they may not perform optimally for specific domains or languages with limited training data. In such cases, n-gram-based methods like BLEU or ROUGE, which focus on surface-level matching, may be more suitable and effective.

In summary, the advantage of sentence\_transformers lies in capturing semantic similarity and contextual information, while the disadvantage is the increased computational requirements and potential limitations in domain or language-specific performance.

**Q2: Computing gender bias in translation (35 points)  
Goal: Inspect accuracy of pronoun translation**

1. Using the dataset in Q1
   1. Extract each of the pronouns in order for the Gold sentence.
   2. Extract each of the pronouns in order for the Translated sentence
   3. Share outputs from (i) and (ii) for three pairs of sentences, e.g., [‘they’, ‘she’, ‘them’] for Gold and [‘he’, ‘she’, ‘him’] for Translated.
2. Using the output in (a), compute the proportion of gold and translated sentences have the same number of pronouns to three decimals of precision.

0.808

1. For sentences in (b) where both sentences have the same number of pronouns, compute co-occurrence (conditional count or conditional probability) statistics between Gold pronouns and Translated pronouns.
   1. For each sentence, compute the proportion of pronouns that match identically *at each position* (for all positions in these two equal-length lists) and store these proportions in a list. Tell us the mean of these proportions to three decimals of precision.

0.816

* 1. Consider the data in 2c.i. Do you see whether errors are more likely to occur when the source pronoun is masculine (“he”, “him”, etc.), feminine (“she”, “her”, etc.) or “they” (“they”, “them”, “theirs”, etc.)? Report the relative proportion of correct translations pronoun class (masculine, feminine, they) and describe the pattern you observe.

Proportions of correct translations:

masculine : 0.962

feminine : 0.227

they : 1.0

1. Name a statistical test that you could use to determine whether errors are higher for some pronouns over others.

One statistical test that can be used to determine whether errors are higher for some pronouns over others is the chi-square test of independence.

The chi-square test of independence can analyze whether there is a significant association between two categorical variables, in this case, the pronouns and the translation errors. By comparing the observed frequencies of errors for each pronoun category with the expected frequencies (assuming independence), the test can determine if there is a statistically significant difference in error rates across different pronouns.

This test can help assess whether errors are more likely to occur for certain pronouns compared to others and provide evidence of any significant associations. However, it's important to ensure that the test assumptions are met, such as having independent observations and an adequate sample size for each pronoun category.

1. **Bonus (2 points)**: Compute the statistic you identify in 2.d

Chi-square Statistic: 1.835204081632653

P-value: 0.39947582017908656

1. **Bonus (2 points)**: Describe one potential problem with using the alignment assumption in 2c. How would this problem make it difficult to assess pronoun translation accuracy?

The one-to-one mapping between pronouns in the gold and translated phrases presents a potential problem with depending on the alignment assumption in mean proportions. This presumption is based on the idea that every pronoun in the source sentence has an exact equivalent in the target sentence. It can be difficult to align pronouns between languages effectively in practice since pronoun usage and translation choices might be complex and context-dependent.

It becomes difficult to evaluate pronoun translation accuracy effectively if there are instances where the alignment assumption is broken, such as when numerous pronouns in the gold text are translated to a single pronoun in the translated sentence or vice versa. It is challenging to analyze and evaluate the precision of pronoun translations across various phrases due to the mismatched alignment, which might result in inflated or deflated proportions.

Furthermore, the alignment assumption does not take into consideration, the differences in sentence structure and word order, which can make determining the accuracy of pronoun translation much more difficult. Direct alignment of pronouns between the source and target sentences might be difficult since various languages may have distinct grammatical rules and word order patterns.

**Q3: Short answer question. (30 points)** Given that machine translation models encode social stereotypes in the statistics that they know, describe a way we might build a machine translation system to minimize the effects of these stereotypes. This should be a solution that either involves using (a) different data and/or (b) different algorithms.

**Bonus 1. (10 points)** Rerun the experiments in Q1 and W2 above using provided data generated with intermediate languages with grammatical gender. We have created two small corpora of parallel text that was generated by translating into and out of (1) French and (2) Hebrew. French has a grammatical gender distinction similar to English for people (masculine and feminine singular pronouns), but the grammar also applies gender to all nouns, so both “the table” and “the grandmother” are feminine. Hebrew has a similar distinction as French, but a very different alphabet whose written words are highly ambiguous out of context. How does translating into a language with grammatical gender differ compared to the Finnish translation system in terms of performance measures in Question 1-f and Question 2-c? Summarize your results.

**Bonus 2. (10 points).** Using provided code (or your own code) for round-trip machine translation, rerun the experiment in 2-c using a language that you know or have experience with. Does this language have grammatical gender? Does the presence or absence of grammatical gender in this round-trip task look similar or different to the results Question 1-f and Question 2-c? Do you think any other factors might affect the quality of backtranslations. Summarize your results.