# Confidence Scoring and WOVEncoding

Daniel Gallagher

February 2022

## 1 Confidence Scoring

Confidence scoring is the conversion of raw Shapley values into a Shapley-based score between 0 and 100, in which a higher value means a greater confidence in a link between two tokens. This takes the form a sparse floating-point matrix.

The matrix is of size n x m, where n represents the output tokens (German) and m represents the input tokens (English). Therefore, each row represents a particular output token. A row contains all confidence scores for the input tokens on a particular output token.

### 1.1 Why do we need to adjust the Shapley values?

In order to determine a connection between and input and output token, we must compare all Shapley values for our inputs on a particular output. This however fails to take into account the Shapley values of these same inputs on all others output tokens. An input token may have a higher Shapley value than another for one output token, yet this value may be small compared to its score on other output tokens. This leads to the need for contextualised scoring, where a given input token's effect on all *other* output tokens is taken into account. This leads to a greater confidence in connections made.

### 1.2 Normalisation

These adjusted Shapley values are now a stronger representation of links between input and output tokens, however they are not easily interpretable. Given that 'confidence' naturally lends itself to a percentage, the values are normalised into the range [0,100]. This is carried out with *min-max normalisation.*

### 1.3 Dealing with Negative Shapley Values

This leaves the case of negative values. These are useful in many tasks for the SHAP approach to explainability, such as in binary sentiment analysis. For instance, negative Shapley values can be interpreted as words contributing to a negative sentiment, while positive contributing to a positive sentiment.

In the case of WOVEN, these values are less useful. They represent the input tokens which are making a negative contribution to an output token, whereas We want to know which input tokens are contributing the most to given output tokens. Thus, negative Shapley values represent something which we are simply not interested in. We will therefore separate these negative values out from the positive using a *negative value identifier*, and henceforth refer to them as the *irrelevant* values.

**Algorithm 1** Confidence Scoring

```
for $iteration = 1, 2, \ldots n$ do
    $shapVals = shap.columns[iteration]$
    for $val = sV_1, sV_2, \ldots, sV_m$ do
        if $val > 0$ then
            $influences := shap.rows[iteration]$
            Remove from influences values less than 0
            Remove $val$ from influences
            if $influences.size > 0$ then
                $averageInfluence = mean(influences)$
            else
                $averageInfluence = 0$
            end if
            $val := val - averageInfluence$
        else
            Store $val$ as an irrelevant value
        end if
    end for
end for
Store all irrelevant adjusted Shapley values as -1
Normalise relevant adjusted Shapley values between 0 and 100
```

# 2 Word-Order Variation Encoding

Word-Order Variation Encoding (WOVEncoding) represents our final encoding from input to output. This is a sparse binary matrix which identifies the connections that we are confident in between input and output. These encodings use the confidence scoring matrix created in the previous step in the pipeline.

The WOVEncoding matrix is of the same size as the confidence scoring matrix. Each row represents a particular output token, and contains all connections made with the input tokens.

## 2.1 Hyper-parameters

There are two hyper-parameters which are used to control whether or not a confidence score is high enough to merit a connection. The first parameter is

$\alpha$, which is the minimum score required to be present in each row for at least one connection to be made. This ensures we do not make connections where there should be none, such as with modal particles which often do not translate directly from German into English.

The second hyper-parameter is $\beta$, which determines whether more than one connection should be made. It is the maximum distance that other confidence scores can be from the highest in that row to make a connection.

These two parameters ensure that: 1. In most cases, at least one connection is made in a given row. 2. In some cases, two or more connections are made in a given row. 3. In few cases, no connections at all are made in a given row.

**Algorithm 2** WOVEncoding

```
$\alpha$ := Threshold that must be reached by at least one confidence score to make an encoding for that row
$\beta$ := Distance from highest confidence score in that row to make more than one connection
for $iteration = 1, 2, \dots m$ do
    $confidenceScores$ := the adjusted Shapley values
    for $row = r_1, r_2, \dots, r_n$ do
        $max$ := the highest confidence in that row
        if $max > \alpha$ then
            $lowerBound := max - \beta$
            if $lowerBound < 0$ then
                $lowerBound := 0$
            end if
            Set all values equal to 1 which are greater than or equal to the lower bound, and all others to 0
        end if
    end for
end for
```