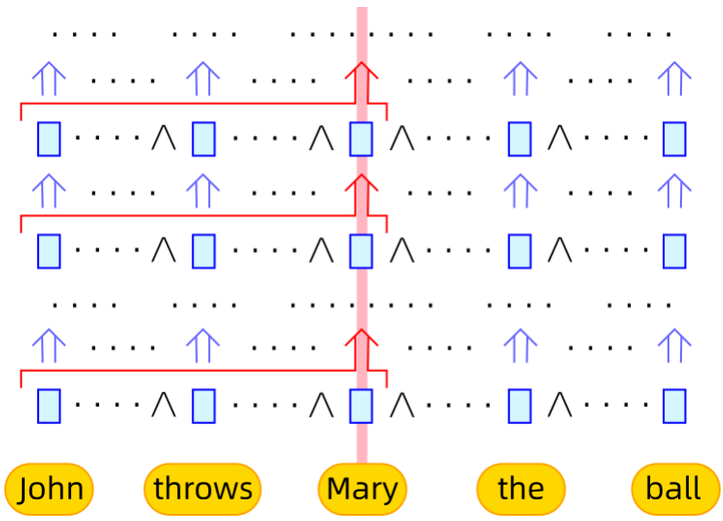
Further Interpreting the Transformer

This is an illustration of how the Transformer processes tokens:



(1)

- The Transformer is basically a neural network function or mapping that accepts a **sequence** token by token, and outputs a new token per each input token. The output of each token is equivalent to looking up a **logic rule base** with premises drawn from the input tokens.
- Each pass of Self-Attention has an **axis** (thick red line), where a **pivot** token is combined with other tokens via super-position.
  Why do pivots exist? The pivot is the **latest** token being added to the "state" of the system, ie, the sequence of tokens up to time t. The rest of the state has not changed except for this "delta" token. Therefore, the logical conclusions that were drawn up to time $t-1$ has already been outputted. What we need to do is to output the new conclusions that is now deducible from the newest token and the rest of the state. This situation is completely analogous to "rule-based systems" in the old days, eg. the SOAR architecture, where tokens are called WMEs (working memory elements).
- Each token seems to be the "summary" of the incoming tokens up to time t. Such a summary seems to be a distributive representation made of superpositions of values $V = W^V X$, but due to the action of *softmax*, some tokens may be **selected** over others. This may have the effect of **disentangled representations** that are closer to symbolic logic.
- We want to see if Self-Attention can perform "variable substitution" in logic rules. This requires extracting some vector components of a token and copying them to the next layer. But this operation varies from rule to rule.