A Data and Setting

A.1 Model architecture

For machine translation models, we follow the setup of the Transformer base model (Vaswani et al., 2017). More precisely, the number of layers in the encoder and in the decoder is $N = 6$. We use $h = 8$ parallel attention layers, i.e. heads. The dimensionality of input and output is $d_{model} = 512$, and the inner-layer of the feed-forward networks has the dimensionality $d_{ff} = 2048$. For language models, we use only the encoder of the model (with the same hyper-parameters).

A.2 Training

Sentences were encoded using byte-pair encoding (Sennrich et al., 2016), with source and target vocabularies of about 32000 tokens. Source vocabulary is the same for all tasks. Minibatch size is set to approximately 15000 source tokens. Training examples were batched together by approximate sequence length. Each training batch contained a set of translation pairs, approximately 15000 source tokens each. The optimizer and learning rate schedule we use are the same as in Vaswani et al. (2017). Since using a large number of training steps was reported to be important for the MLM objective,we follow Devlin et al. (2018) and train MLM for 1 million training steps and other models till convergence.

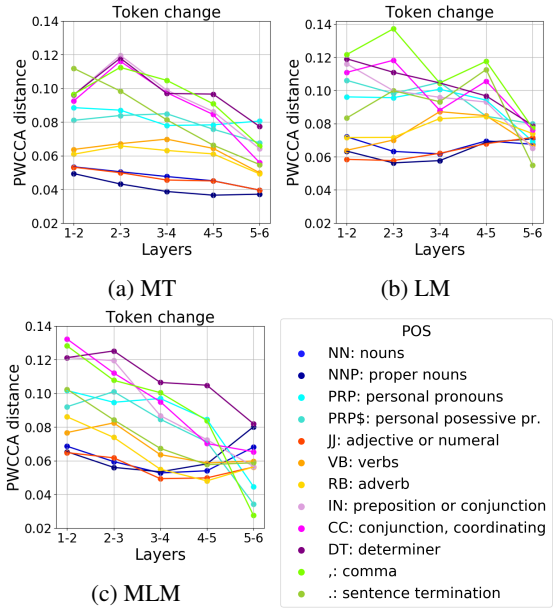
B Fine-grained analysis of change and influence: varying PoS

Figure 1 shows the amount of change for different parts of speech, Figure 2 – the amount of influence for different parts of speech[1]. Generally, the patterns are similar to the ones for frequency groups: parts of speech with frequent tokens (preposition, conjunction, etc.) change more and influence less.

C What does a layer represent?

C.1 Preserving token identity: experimental setup

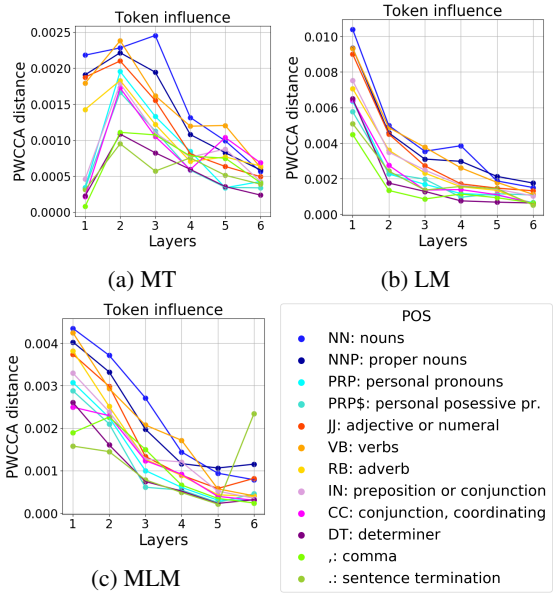
In this section, we want to check to what extent a model confuses representations of different tokens. For each of the selected tokens described above ("main" tokens), we pick tokens which potentially could be confused with the token under consideration ("contrastive" tokens). In our setting, contrastive tokens are the top 10 closest tokens in the embedding space (separately for each model). For example, tokens "is", "are", "was", "were" are close in the embedding space of each of the models. Then for each main token we gather representations of 9000 different occurrences of its contrastive tokens, and add them to the 1000 states of the main token. All in all, we get 200 groups of representations; each group contains representations of 1000 occurrences of the main token and 9000 of the contrastive ones (at each layer). For representations of the main token in a group, we



(a) MT (b) LM

(c) MLM

Figure 1: Token change vs its part of speech.



(a) MT (b) LM

(c) MLM

Figure 2: Token influence vs its part of speech.

[1]We use the part-of-speech tagger from Stanford CoreNLP (Manning et al., 2014).

measure the average percentage of representations of the same token for top closest among 10000 representations.