Attention Regularization Improves Counter Narrative Generation

HS

Any migrant can become a citizen even if he is a criminal. This is how you destroy the welfare state.

Regularized HS

Any migrant can become a citizen even if he is a criminal. This is how you destroy the welfare state.

CN

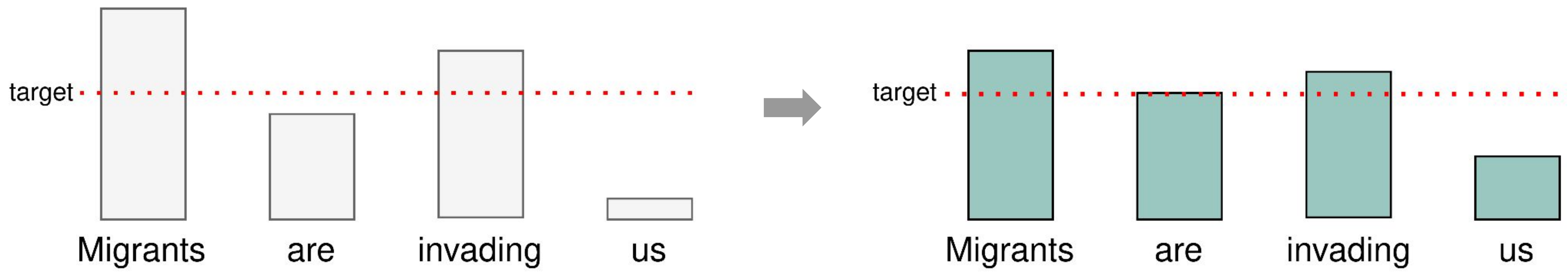
I don't understand why you think this way about migrants, they are just people trying to make a better life for themselves.

Regularized CN

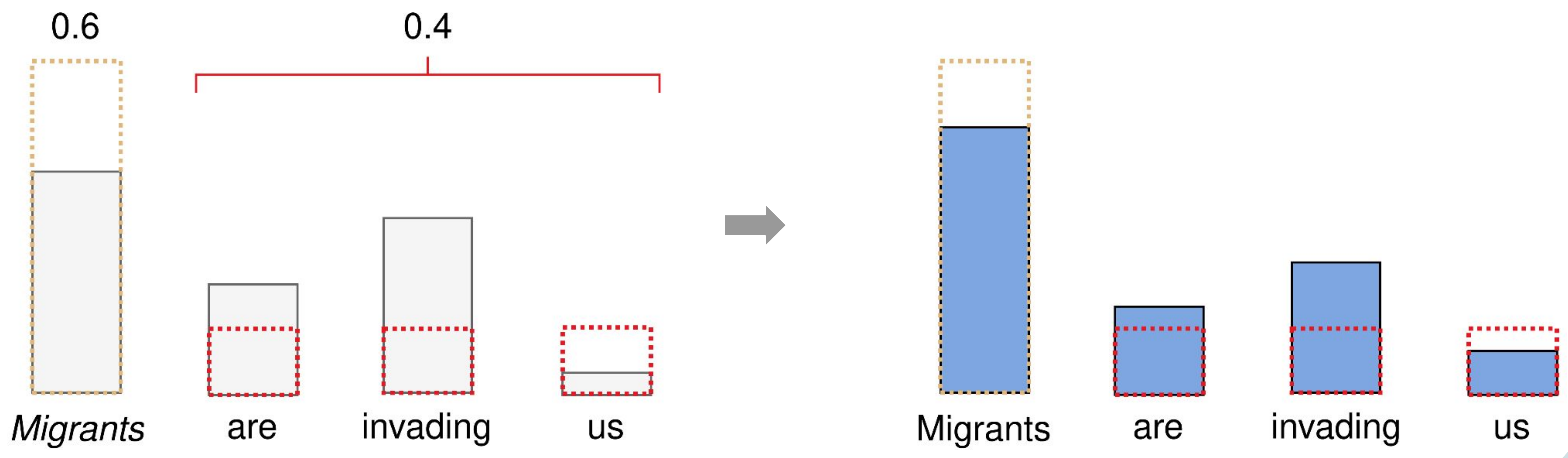
The right to live and work according to one's beliefs is guaranteed by the European Convention on Human Rights, which also includes the right to respect for private and family life.

We test two attention-based regularization techniques on GPT-2:

Entropy-based AR: aims to maximize each token's attention entropy.



Kullback-Leibler AR: particular attention is posed to relevant tokens.



Results

📈 **KLAR** obtains the highest overlap scores with gold CNs

🔍 **EAR** reaches the highest human evaluated specificity in most tested cases

💬 Regularization however brings a higher repetitiveness

✅ Results obtained by regularization are robust to Leave One Target Out setup

"Weigh Your Own Words: Improving Hate Speech Counter Narrative Generation via Attention Regularization" H. Bonaldi, G. Attanasio, D. Nozza, M. Guerini