Μηχανές που γράφουν
Εντοπισμός κειμένων από μοντέλα τεχνητής νοημοσύνης και οιπαιδαγωγικές τους προεκτάσεις

Γιώργος Μικρός
Department of Middle Eastern Studies,
College of Humanities and Social Sciences (CHSS), Hamad Bin Khalifa University (HBKU)
gmikros@hbku.edu.qa

Περίληψη

Κατά τη διάρκεια των τελευταίων ετών, η εμφάνιση των Μεγάλων Γλωσσικών Μοντέλων (Large Language Models -LLMs) έχει φέρει επανάσταση στον κλάδο της Επεξεργασίας Φυσικής Γλώσσας, αλλά και τον ευρύτερο τομέα της Τεχνητής Νοημοσύνης. Τα συγκεκριμένα μοντέλα μπορούν πλέον να αντιμετωπίσουν με επιτυχία τις περισσότερες εργασίες Κατανόησης Φυσικής Γλώσσας (Natural Language Understanding - NLU) και Γλωσσικής Παραγωγής (Language Generation - LG) προσφέροντας κείμενα υψηλής ποιότητας με συνοχή και θεματική εστίαση (Brown et al., 2020). Στις περισσότερες περιπτώσεις, τα κείμενα αυτά είναι τόσο καλογραμμένα που δεν μπορείς να τα διακρίνεις από τα κείμενα ενός ανθρώπου. Μέχρι τώρα, οι περισσότερες προσεγγίσεις για την ανίχνευση του αν ένα κείμενο έχει γραφεί από ένα γλωσσικό μοντέλο ή από άνθρωπο αποτυγχάνουν ή, στην καλύτερη περίπτωση, δίνουν αντιφατικά αποτελέσματα που δεν είναι αρκετά αξιόπιστα για να χρησιμοποιηθούν σε πραγματικές εφαρμογές εντοπισμού γραφής Τεχνητής Νοημοσύνης (Fröhling &Zubiaga, 2021· Solaiman et al., 2019· Varshney et al., 2020). Ο εντοπισμός γραφής των Μεγάλων Γλωσσικών Μοντέλων έχει γίνει ακόμη πιο δύσκολος με την κυκλοφορία του ChatGPT, του νεότερου LLM από την εταιρεία Open AI. Το ChatGPT βασίζεται στην οικογένεια των γλωσσικών μοντέλων GPT- 3 και μπορεί να δημιουργήσει ακόμα πιο ανθρώπινο και συνεκτικό κείμενο από τους προκατόχους του. Αυτό έχει οδηγήσει σε μια παγκόσμια ανησυχία σχετικά με την πιθανότητα κακόβουλης χρήσης αυτών των μοντέλων όπως π.χ., η δημιουργία και διασπορά ψευδών ειδήσεων, η πλαστοπροσωπία ατόμων στο διαδίκτυο, η στοχευμένη χειραγώγηση χρηστών μέσα από τα κοινωνικά δίκτυα κ.ά.

Στην ομιλία αυτή θα παρουσιάσουμε τις βασικές αρχιτεκτονικές των Μεγάλων Γλωσσικών Μοντέλων με έμφαση στο ChatGPT, θα δείξουμε τι μπορούν και δεν

μπορούν να επιτύχουν μέχρι σήμερα. Επίσης θα γνωρίσουμε ορισμένες ερευνητικές προσπάθειες για τον αυτόματο εντοπισμό γραφής Μεγάλων Γλωσσικών Μοντέλων μέσα από τις τεχνολογίες μηχανικής μάθησης. Θα κλείσουμε τη συγκεκριμένη παρουσίαση συζητώντας τις προεκτάσεις αυτών των τεχνολογιών στην εκπαίδευση καθώς και πιθανές πρακτικές δραστηριότητες που οι εκπαιδευτικοί θα μπορούσαν να ενσωματώσουν στην καθημερινή διδακτική πράξη ώστε οι επιπτώσεις αυτών των τεχνολογιών να μην διαβρώσουν την ποιότητα και την εγκυρότητα της εκπαιδευτικής διαδικασίας.

Βιβλιογραφικές Αναφορές

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., . . . & Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems 33 (NeurIPS 2020), 33*, 1877-1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Fröhling, L., & Zubiaga, A. (2021). Feature-based detection of automated language models: Tackling GPT-2, GPT-3 and Grover. *PeerJ Computer Science, 7*, e443. https://doi.org/10.7717/peerj- cs.443

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J. W., Kreps, S., & others (2019). Release strategies and the social impacts of language models. *arXiv preprint arXiv:*1908.09203.

Varshney, L. R., Keskar, N. S., & Socher, R. (2020). Limits of detecting text generated by large-scalelanguage models. In *2020 Information Theory and Applications Workshop (ITA)* (pp. 1-5). https://doi.org/10.1109/ITA50056.2020.9245012