# Review 130: SeamlessM4T—Massively Multilingual & Multimodal Machine Translation, 27.08.23 https://ai.meta.com/research/publications/seamless-m4t

**Paper: https://arxiv.org/abs/2308.11596v3**

היום ב-#shorthebrewpapereviews נסקור מאמר מרשים מבית היוצר של מטה שיצא לפני כמה ימים. המאמר מציע מודל מולטימודלי שתומך בשני סוגים של דאטה בלבד: אותות קול (speech) ושפה. זה לא נשמע מרשים מדי עד שמבינים מה המודל הזה מסוגל לעשות עם שני סוגי הדאטה האלו. למעשה המודל מאפשר לבצע 4 פעולות: תרגום של אות קול משפת מקורה לשפה אחרת, תרגום של אות קול לטקסט בשפה אחרת, הפיכה של אות קול לטקסט בשפה אחרת ותרגום טקסטואלי רגיל.   
  
זה גם לא נשמע מרשים במיוחד בתקופתנו העמוסה במודלים עוצמתיים שיוצאים כמעט כל יום עד שמגלים ה- SeamlessM4T יודע לבצע את הפעולות האלו בלא פחות מ-200 שפות שונות כולל שפות די לא נפוצות כמו גאורגית וליטאית. כמו שכבר אמרנו בסקירות הקודמות הדבר החשוב ביותר במודלי מולטימודליים הוא מיפוי קוהרנטי של סוגי הדאטה השונים לאותו מרחב שיכון (embedding space).   
  
הקוהרנטיות כאן פירושה שפיסות דאטה מסוגים עם משמעות דומה יהיו קרובים במרחב הזה ואלו שמשמעותם שונים יהיו רחוקים זה מזה. מכיוון שהמבנה הפנימי של אות קול הוא מאוד שונה מטקסט הוא צריך לעבור עיבוד מקדים (preprocessing) לפני שמכניסים אותו לרשת הממפה אותו למחרב הזה.   
  
מכיוון שגלי קול בשפות שונות מבנה מאוד שונה בשלב הראשון מזהים את השפה (כאשר הקלט מכיל כמה שפות מפרקים אותו לסגמנטים שכל אחד הוא בשפה שונה). לאחר מכן לוקחים אות דיבור בכל שפה ועושים מה שנקרא oversegmegmentation שמשעומתה מספר חלוקות של האות לכמה סגמנטים שונים. ייצוגים אלו נכנסים לאנקודר של אותות קול שמאומן להפיק וקטור שיכון של אות קול.   
  
המאמר גם מאמן אנקודר לטקסט בצורה די סטנדרטית (אך לשפות מרובות). לאחר מכן מאמנים מודל הממפה שיכונים אלו למרחב ייצוג משותף ולאחר מכן מאמנים דקודר ההופך את ייצוג הזה לטקסט. בשלב האחרון מאמנים מודל נוסף ההופך את הטקסט הזה לאות קול. יש הרבה פרטים מעניינים על תהליך האימון שלא ניתן לכסותם בסקירה הקצרה הזו – ממליץ להעיף מבט.