

Beágyazási modellek értékelése magyar nyelvű információ-visszakereséshez

Antal Margit, Csiszér Rolland-Zsombor

Sapientia EMTE, Marosvásárhely, Románia

manyi@ms.sapientia.ro, csiszer.rolland@student.ms.sapientia.ro

A beágyazási modellek a modern természetes nyelvfeldolgozás (NLP) alapvető építőelemeivé váltak, ugyanakkor teljesítményük a morfológiailag gazdag, alacsony erőforrású nyelvekben – mint például a magyar – mindeddig kevésbé feltárt. Jelen tanulmányban a legkorszerűbb beágyazási modellek szisztematikus értékelését mutatjuk be magyar nyelvű kérdés–válasz visszakeresési feladaton. Két, egymást kiegészítő adathalmazon végeztünk méréseket: (i) egy doménspecifikus korpuszon, amely vállalati dokumentációból származik, és tematikus egységekre bontva, ember által ellenőrzött kérdés–válasz párokkal került előfeldolgozásra; valamint (ii) a nyilvánosan elérhető HuRTE benchmarkon. A FAISS vektoralapú adatbázist alkalmazva nyolc többnyelvű beágyazási modellt hasonlítottunk össze. A teljesítményt Mean Reciprocal Rank (MRR) és Recall@k mutatók segítségével értékeljük. Az eredmények jelentős varianciát mutatnak a modellek és az adathalmazok között, különösen a doménspecifikus és az általános célú visszakeresési feladatok esetében. Az értékelést hibaelemzéssel egészítjük ki, amely feltárja a magyar morfológia és szóösszetétel okozta nehézségeket, valamint megvitatjuk a hatékonysági kompromisszumokat.

Kulcsszavak: beágyazások, visszakeresés, magyar nyelv

References

- [1] Antal, M., Buza, K. (2025) Evaluating Open-Source LLMs in RAG Systems: A Benchmark on Diploma Theses Abstracts Using Ragas. *Acta Univ. Sapientiae Inform.* 17, 5 (2025). <https://doi.org/10.1007/s44427-025-00006-3>
- [2] Ligeti-Nagy, N., Ferenczi, G., Héja, E., Jelencsik-Mátyus, K., Laki, L. J., Vadász, N., Yang, Z. Gy. and Váradi, T. (2022) HuLU: magyar nyelvű benchmark adatbázis kiépítése a neurális nyelvmodellek kiértékelése céljából [HuLU: Hungarian benchmark dataset to evaluate neural language models]. In: Berend, G., Gosztolya, G. and Vincze, V. (eds), XVIII. Magyar Számítógépes Nyelvészeti Konferencia. Szeged, Szegedi Tudományegyetem, Informatikai Intézet. 431–446.