1. félévi beszámoló

Asztalos Bogdán (*abogdan@caesar.elte.hu*)

Statisztikus Fizika, Biológiai Fizika és Kvantumrendszerek Fizikája
PhD program

Témavezető: Pollner Péter

A dolgozat címe: Erősen kölcsönható, történetfüggő, nem-egyensúlyi komplex rendszerek vizsgálata

Bevezetés

A természettudományok minden területén szükség van olyan rendszerek tanulmányozására, amelyek rendkívül sok, egymással kölcsönható elemből állnak és összetett szerkezetet alkotnak. Ezen rendszerekre általában az jellemző, hogy viselkedésük csak nagyszámú változó segítségével írható le, amelyek egymástól bonyolult módon függhetnek, így nem vizsgálhatóak egyszerű modellekkel. Ehelyett, egy-egy jelenség tanulmányozásakor a változók kollektív dinamikájának statisztikus jellemzőit figyelhetjük meg, és az így kapható makroszkopikus jellemzőkből következtethetünk a dinamikát leíró mikroszkopikus szabályokra, illetve prediktív modelleket állíthatunk fel további kollektív statisztikus jellemzőkre vonatkozóan. A célom, hogy doktori tanulmányaim során egyrészt ilyen statisztikus jellemzőkre vonatkozó törvényszerűségeket tárjak fel elméleti statisztikus fizikai vizsgálatok segítségével, másrészt pedig, hogy konkrét rendszereket vizsgálva ismerjem meg az őket mozgató mikroszkopikus szabályokat.

Egy korábban megkezdett kutatási projekt nyomvonalán haladva, ebben a félévben főleg a természetes nyelvek időfejlődésével foglalkoztam, azon belül is a szavak időbeli jelentésváltozását vizsgáltam az utóbbi években széles körben elterjedt Word2vec szóbeágyazási algoritmus segítségével. A 2010-es években a mesterséges neurális hálók alkalmazásának elterjedése a természetes nyelvek tudományos tanulmányozását is jelentősen befolyásolta: megjelentek az olyan szóbeágyazási módszerek[1], amelyek a beágyazást nem valamilyen szabály vagy számolás, hanem gépi tanulás alapján végzik. Ezek közül a legnépszerűbb a Google által fejlesztett, 2013-ban publikált Word2vec algoritmus [1, 2], amely a tapasztalat alapján igen hatékony reprezentációját adja a nyelvnek. Különböző időpontokból származó adatok alapján végezve szóbeágyazást a szavakra úgy tekinthetünk, mint a geometriai térben bolyongó részecskék halmazára, így a viselkedésük a statisztikus fizika eszközeivel tanulmányozható.

Az aktuális félévben elvégzett kutatás

A félév során két külön területen foglalkoztam a szavak jelentésváltozásával. Az egyik terület a korábbi kutatási témám folytatása volt, ahol általánosságban vizsgáltam a szavak mozgásának dinamikáját a beágyazási térben. A másik pedig az itt megszerzett technikai tapasztalatoknak egy konkrét alkalmazása, ahol általános nyelvi források helyett internetes közösségi felületekről gyűjtött adatok alapján végeztem beágyazást, és a covidjárványhoz köthető, külső események hatására történő szójelentés-változások megfigyelését és leírását végeztem.

[1] A szóbeágyazás egy olyan technika, amely egy természetes nyelv szavait az n-dimenziós térbe képezi le úgy, hogy a szavak közötti jelentésviszonyok a geometriai távolságokkal legyenek kifejezhetők.

Szavak jelentésváltozásának dinamikája

Ezen kutatás során feltett fő kérdés az volt, hogy ha a Word2vec beágyazó algoritmussal egymás utáni időperiódusokból származó adatok alapján készítünk beágyazást, akkor milyen megfigyeléseket tehetünk a beágyazott szavaknak mint bolyongó pontoknak a mozgására. A beágyazások elkészítéséhez egy stanfordi kutatócsoport által publikált kódot használtam [3], az eredmények feldolgozását pedig csoportos megbeszélések során egyeztett ötletek alapján saját magam végeztem.

A vizsgálatok alapján talált legfontosabb megfigyelhető eredmény a szavak mozgásának korreláltsága, ugyanis azt tapasztaltuk, hogy a beágyazási térben a szavak szubdiffúzív viselkedést mutatnak. Mivel az egymástól függetlenül, időben korrelálatlan, véletlenszerű mozgást végző részecskék halmaza diffúziót mutatna, ésszerű lenne az a feltételezés, hogy a szavak esetében is ezt várjuk. Ennek fényében ez az eredmény nagyon érdekes, amely további új kérdéseket vethet fel a nyelvi változások vizsgálatának területén.

A jelenség tanulmányozásának fő kihívása az, hogy a látott szubdiffúzió lehetséges okai közül kizárjunk minden technikai vagy adatgyűjtési effektust, hogy így teljes bizonyossággal kijelenthessük, hogy a megfigyelések a nyelv viselkedéséhez köthető. Ehhez megvizsgáltam a nyers nyelvészeti adatokat sűrítő együttelőrdulási mátrixnak a mátrixtérben való mozgását, és a feldolgozás lépéseit fokozatosan végrehajtva figyeltem, hogy ez hogyan viszonyul a beágyazási térben látott kollektív mozgáshoz. Amellett, hogy ezek során olyan nyelvi törvényszerűségek voltak felfedezhetők mint például a szókapcsolatokra vonatkozó Zipf-törvény, az is megfigyelhető volt, hogy a beágyazási eljárás során a nagyobb frekvenciájú szavak veszítenek matematikailag domináns szerepükből, és a kolletív viselkedést sokkal inkább befolyásolják a ritkábban előforduló kifejezések.

Jelen ismeretünk szerint elég erős argumentum van annak belátására, hogy a beágyazási térben látott szubdiffúzió valóban nyelvi jelenség, és a szavak viselkedéséről hordoz információt. Ezen eredmény publikálása folyamatban van, reményünk szerint a következő félév során kéziratot is benyújthatunk.

A covidjárvány hatása az online szóhasználatra

Az előző projektben használt beágyazási folyamat nem csak általános nyelvészeti vizsgálatra használható, hiszen célzottan szűrt adatok alapján konkrétabb kérdések kapcsán is lehet alkalmazni. Ez a motiváció állt amögött, hogy ebben a félévben egy olyan nyelvi adathalmaz alapján végzett beágyazást is tanulmányoztunk, amely online közösségi felületekről lett legyűjtve kimondottan a covidjárvánnyal kapcsolatos keresőszavak alkalmazásával. Az adatok a Nemzeti Közszolgálati Egyetem egy kutatócsoportjával való együttműködésből származnak. 2019. július és 2021. június közötti időszakból származó adatokból minden hónaphoz külön-külön beágyazást készítettem, és olyan viselkedési mintákat kerestem, amelyeken megfigyelhető volt, hogy a társadalmi események észrevehető módon befolyásolják egy-egy szó jelentését a közösségi felületeken.

Főként az oltással kapcsolatos kifejezéseknél és a vakcinák neveinél volt nagyon feltűnő az a jelenség, hogy a járvány különböző szakaszainál hirtelen nagy jelentésváltozáson mennek át. Ilyen volt például a *Pfizer* szó, amely a covid megjelenése előtt vegyészettel és általános gyógyszerészettel kapcsolatos jelentést mutatott, de 2020. tavaszától eltolódott a betegség és a járvány témaköréhez, majd az év második felétől egyértelműen az oltóanyaggal azonosították az emberek: 2020. október és 2021. februárja között a legközelebbi szomszédja az *experiment* szó volt, majd 2021. márciusától az injekció informális megfelelői, a *shot* illetve a *jab*.

Ezen kívül kiszámoltam a szavak havi jelentésváltozását jellemző euklideszi lépésnagyságot is különböző szavakra. Általában jól megfigyelhető volt, hogy mely szavak jelentésében mikor történik hirtelen változás, és ez általában elég erősen korrelált a járvány okozta társadalmi eseményekkel. Így például a *mask* szó 2020. januárjában és februárjában, a *lockdown* szó

2020. márciusában, a *vaccine* szó pedig 2021. januárjában változot meg kiemelkedő mértékkel. Reményeink szerint az ehhez hasonló eljárások lehetőséget biztosíthatnak arra, hogy a vizsgálódás iránya megfordítható legyen: a kutatás során társadalmi jelenségek nyomait kerestem a közösség felületek alapján készített beágyazási adatokban, de elméletben ezek a megfigyelések arra is alkalmasak lehetnek, hogy pusztán a közösségi felületeket vizsgálva észrevegyünk és beazonosítsuk olyan társadalmi hatások kiindulópontját mint a lezárások miatti tömegpánik, a vakcinák iránti érdeklődés vagy éppen az oltásszkepticizmus.

Publikációk

Mindkét fent részletezett téma eredményeit tervezzük publikálni, a cikkek kéziratai jelenleg készülőben vannak. Mivel az utóbbi kutatás az előbbi módszereit használja, terveink szerint először ezt tesszük majd nyilvánossá, ezzel hivatkozási alapot teremtve a későbbi cikknek. Jelen állás szerint mindkét publikáció el fog tudni készülni a következő félév során.

Tanulmányi tevékenység

A félév során az alábbi tárgyakat végeztem el:

- Extrémek, rekordok és sorrend-statisztikák a természetben (FIZ/3/075E)
- Fejlődésbiológiai mechanizmusok kvantitatív modelljei (FIZ/3/056E)

Oktatási tevékenység

A félév során a Fizika BSc szakos hallgatóknak meghirdetett *Modern fizika laboratórium* (fizlab3f19la / ff1c4s13) című tárgy lebonyolításában vettem részt gyakorlatvezetéssel (6 alkalommal vezettem a 3+1 órás gyakorlatot) és a *kvantumradír* méréshez készült jegyzőkönyvek (összesen 21 darab) értékelésével.

Szakmai közéleti tevékenység

A doktori tanulmányaim mellett igyekszem aktívan részt venni a középiskolások természettudományos tehetséggondozásában is. Ennek keretében ebben a félévben is havi szinten javítottam a Középiskolai Matematikai és Fizikai Lapok (Kömal) által kiírt levelező pontversenyre beérkező feladatokat, illetve részt vettem a Dürer verseny fizika kategóriáinak (F és F+ kategória) helyi fordulós feladatsorának összeállításában.

Elismerések

A 2021/2022-es tanév során támogatásban részesülök az Új Nemzeti Kiválóságok Program keretében (ÚNKP-21-3-I-ELTE-228).

Hivatkozások

[1] Tomas Mikolov, et al. *Distributed representations of words and phrases and their compositionality.* Advances in neural information processing systems. 2013.

[2] Tomas Mikolov, et al. *Efficient estimation of word representations in vector space.* arXiv preprint arXiv:1301.3781 (2013).

[3] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. *Diachronic word embeddings reveal statistical laws of semantic change.* arXiv preprint arXiv:1605.09096 (2016).