

4. 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

Az előző három félévben végzett kutatások összegzése

Szavak jelentésváltozásának dinamikája

Az alapkérdés, amit feltettünk az volt, hogy milyen törvényszerűségek figyelhetők meg a nyelvi változások során. Ennek matematikai vizsgálatában segítségünkre volt a Word2vec szóbeágyazó algoritmus, amely segítségével a szavakat el lehetett helyezni a jelentésük alapján egy sokdimenziós beágyazási térben. Időcímkékkel ellátott nyelvi adatokból kiindulva különböző szóbeágyazásokat készítettünk, amelyekre úgy tekintettünk mint a nyelv különböző állapotairól készült snapshotok, így mérési eljárást kaptunk a szavak „jelentésében” bejárt trajektóriájáról.

A kutatás során egy olyan metodológiát dolgoztunk ki, amellyel különböző idők alapján készített egyedi beágyazások egymás után fűzhetők. Ez használható lehet más jellegű beágyazási módszerek során, illetve egyéb nyelvi jellegű kérdések vizsgálatához (lásd. pl. következő pont). A nyelvi dinamikára azt az eredményt találtuk, hogy a szavak többsége folyamatos mozgásban van (mintegy termikus alapzaj), amely szubdiffúzív jellegű. 5 különböző nyelvre lemérve az anomális diffúziós exponenst, azt kaptuk, hogy ennek értéke szignifikánsan kisebb 1-nél, és értékük 0,2 és 0,5 között mozog.

COVID-járvány hatása az online nyelvhasználatra

Az előző kutatási téma során kidolgozott metodológiai újításokat egy mindennapi élethez közelebb álló téma kapcsán is felhasználtuk. Online felületekről letöltött, járvánnyal kapcsolatos szövegek elemzését végeztük, és ezek alapján készítettünk időfüggő szóbeágyazást a 2019. július és 2021. június közötti időszakra. Azt vizsgáltuk, hogy a szavak térben mozgó pontokként való leírásával milyen konkrét jelentésváltozásokat tudunk detektálni, és hogy ezek mit árulnak el azokról a társadalmi változásokról, amelyek az adott időintervallumban történtek.

Eredményként azt kaptuk, hogy valóban megfigyelhetők olyan átrendeződések a szavak viszonyában, amelyek a járvány ideje alatt társadalmi folyamatokban is megfigyelhető volt (például az öt fő vakcinagyártó cég neve határozottan közel került egymáshoz, vagy mondjuk az *oltás* szó először a *gyógymód* közeléből átvándorolt a *kísérlet*, majd az *adag* közelébe). Ezen a változások kvantitatív jellemzésére a legközelebbi szomszédok által alkotott halmazok közötti szimmetrikus differenciát használtuk, amely segítségével még több covid-jelenséggel kapcsolatos szóváltozásokat tudtunk találni. Reményeink szerint az általunk használt (és a jövőben publikálható módszer) használható lehet szociális jelenségek valós idejű követésére vagy akár előrejelzésére is.

Közösségi médiáról gyűjtött hálózatok elemzése

A CounteR nemzetközi együttműködés keretében olyan szoftvermodul fejlesztésében is részt vettem, amely célja, hogy hálózattudományi eszközökkel segítse az online radikalizálódás folyamatát időben felismerni és megakadályozni. Az előző félévekben ennek kapcsán a feladatom főként a kapott nyers adatok struktúrájának megismerése volt, és azon módszerek megismerése és módszeres dokumentálása, amelyek segítségével az adatokban található hálózat jellegű információk kinyerhetők. Ezen célból először megismerkedtem a Twitter adatgyűjtő API-jával, amely segítségével radikalizálódott és potenciálisan veszélyes felhasználók bejegyzéseiből kiindulva készítettünk hálózatokat. Ezután egy együttműködési partner által készített nyelvi elemző modul kimenetelét tanulmányoztam át abból a szempontból, hogy a bejegyzésekhez, illetve a felhasználóhoz tartozó szavak és névntitások felhasználásával készült hálózatok mennyire lehetnek alkalmasak a céljaink számára.

Az aktuális félévben elvégzett kutatás

Ebben a félévben főként az alábbi két kutatási területen születtek főbb eredmények.

Csoportkeresési algoritmusok elemzése radikalizációs hálózatokon

A CounteR nemzetközi együttműködés keretében olyan szoftvermodul fejlesztésében is részt vettem, amely célja, hogy hálózattudományi eszközökkel segítse az online radikalizálódás folyamatát időben felismerni és megakadályozni. Mivel az online radikalizálódás nagyon sokszor kisebb csoportokba szerveződik, ezért releváns kérdés az, hogy potenciálisan gyanús felhasználóknak már beazonosított csoportját hogyan lehet minél jobban akadályozni az összetartásban. Az együttműködésben részt vevő hatósági szervek ettől azt várják, hogy az így önállóan maradó embereknek a veszélyes irányba való közeledése lelassul vagy akár vissza is fordul.

Ezzel kapcsolatban azt vizsgáltuk, hogy a hálózatok területén korábban is ismert csoportkereső algoritmusok által csoportszerekezet hogyan gyengíthető meg minél hatékonyabban. Két csoportkereső algoritmust vizsgáltunk: a label propagation módszert [1] és a louvain algoritmust [2]. Ezek segítségével jellemeztük a különböző módszerekkel gyengített csoportszerekezetek meglétét, és a módszerek hatékonyságát. Ugyan a vizsgálatok eredménye a hatósági szervekkel való konzorciumi együttműködés miatt egyelőre bizalmas státuszba tartozik, ezért ebben a beszámolóban nem részletezem, de sikeresen rangsoroltuk a módszereket annak függvényében, hogy milyen hamar szünteti meg a meglévő csoportokat egy-egy konkrét radikalizációs hálózatban.

Hierarchikus korrelációk vizsgálata

Ebben a félévben folytattuk a korrerált rendszerek hierarchikus generatív folyamatokkal való leírásának vizsgálatát. Itt főként elméleti számolások segítségével próbáltuk különböző eseteket megtalálni a hierarchikus leágazás segítségével létrejövő korrelált változóknak egy adott minimálmodell segítségével. Mivel már korábban sikerült a korrelációra jellemző N_{eff} effektív szabadsági foknak elnevezett mennyiséget kifejezni a C_N korrelációs mátrix segítségével az alábbi képlet segítségével:

$$N_{\text{eff}}(N) = N + \frac{\ln \det C_N}{\ln(2\pi e) + \ln(2DT)}, \quad (1)$$

ezért már csak a C_N N -függésére próbáltunk fókuszálni. Különböző közelítő módszereket alkalmazva, el lehet jutni két rekurzív kifejezésre, amely közül az egyik a lehetséges C_N esetek

alsó korlátját, a másik a felső korlátját adja meg. Ezen két egyenlet:

$$|C_N| = 2(1 - \tau_N/T) \cdot |C_{N-1}| - (1 - \tau_N/T)^2 \cdot |C_{N-2}| \quad (2)$$

$$|C_N| = 2^N \cdot \prod_{i=2}^N (1 - \tau_N/T) \quad (3)$$

Mivel konkrét τ_N függések esetén mindkét egyenlet megoldható, ezért különböző esetekre kapható konkrét eredmény, amely alapján az analitikus függvények különböző családra oszthatók aszerint a szempont szerint, hogy a hierarchikus leágazás rövidtávú vagy hosszútávú korrelációt produkál a kapott rendszerben.

Publikációk

A teljes képzés során két kézirat készült a kutatások keretében:

- Dr. Palla Gergellyel és Czégel Dániellel közösen adtuk be a „Subdiffusive Semantic Evolution in Indo-European Languages” című cikket a *Scientific Reports* című újsághoz. Ez jelenleg szakmai elbírálás alatt van. A kézirat preprintje elérhető az arXiv szerveren [3].
- Dr. Bányász Péter, a Nemzeti Közszolgálati Egyetem adjunktusának társszerzőségével készült egy cikk „Monitoring the semantic change of Covid-19-related expressions using dynamic word embeddings” címmel, amelyet némi átdolgozás után a *Journal of Communications* lapba tervezzük küldeni.

Emellett előkészületben van egy harmadik kézirat a hierarchikus korrelációk témájából, amelyet néhány hónapon belül az *Entropy* című lapba Dr. Palla Gergellyel és Czégel Dániellel.

Konferenciák

A képzés során az alábbi nemzetközi konferenciákon vettem részt:

- 2022. május 16. és 22. között a *Complex Networks: Theory, Methods and Applications* című tavaszi iskolán a *Lake Como School of Advanced Studies* programsorozat keretében az olaszországi Como-ban.
- A *The 2022 IEEE 2nd Conference on Information Technology and Data Science* című konferencián, ahol 2022. május 17-én online előadást tartottam *Tracking the time evolution of Covid-19 related words via word2vec*.

Utóbbin tartott előadást két magyar nyelvű rendezvényen is előadtam:

- 2022. október 5-én *A hálózattudomány aktuális kérdései* című, az NKE Hálózattudományi Kutatóműhely által szervezett konferencián
- 2022. november 25-én a Statisztikus Fizikai Társaság által rendezett *Statisztikus Fizika Nap* rendezvényen

Tanulmányi tevékenység

A félév során az alábbi tárgyat végeztem el:

- Haladó statisztika és modellezés (FIZ/3/088)

Oktatási tevékenység

A félév során a Fizika BSc. szakos hallgatóknak meghirdetett *Modern fizika laboratórium* (fiz-lab3f191a) című tárgy lebonyolításában vettem részt gyakorlatvezetéssel (5 alkalommal vezettem a 3+1 órás gyakorlatot) és a *kvantumradír*, valamint a *granuláris anyagok* mérésekhez készült jegyzőkönyvek (a beszámoló leadásáig 18 darab) értékelésével.

Szakmai közéleti tevékenység

A doktori tanulmányaim mellett igyekszem aktívan részt venni a középiskolások természettudományos tehetséggondozásában is. Ennek keretében ebben a félévben részt vettem a Baár–Madar Református Gimnázium fizikaszakkörének tartásában, illetve az országos Dürer verseny fizika kategóriáinak (F és F+ kategória) országos forduló feladatsorának összeállításában és a miskolcon rendezett háromnapos döntő lebonyolításában is közreműködtem.

Hivatkozások

- [1] G. Cordasco, L. Gargano, *Community detection via semi-synchronous label propagation algorithms*. 2010 IEEE International Workshop on Business Applications of Social Network Analysis (BASNA) (pp. 1-8).
- [2] V. D. Blondel, et al. *Fast unfolding of communities in large networks*. Journal of statistical mechanics: theory and experiment 2008.10 (2008): P10008.
- [3] B. Asztalos, G. Palla, D. Czégel. *Subdiffusive semantic evolution in Indo-European languages*. arXiv preprint. arXiv:2209.04701