Nettikorpuksen avulla tuotettuja sanavektorimalleja Pokémonien ominaisuuksien kuvaamiseksi

Hämäläinen, Mika

Suomalaisen Kirjallisuuden Seura
2021

Hämäläinen , M , Alnajjar , K & Partanen , N 2021 , Nettikorpuksen avulla tuotettuja sanavektorimalleja Pokémonien ominaisuuksien kuvaamiseksi . julkaisussa S Taina & S þÿ J a n n e (t o i m) , T u r h a n t i e d o n k i r j a T u t k i m u k s i s t a p o i s j ä t e t t y j ä s i v u j a Suomalaisen Kirjallisuuden Seura , Sivut 199-214 .

http://hdl.handle.net/10138/333995

acceptedVersion

Nettikorpuksen avulla tuotettuja sanavektorimalleja Pokémonien ominaisuuksien kuvaamiseksi

Mika Hämäläinen, Khalid Alnajjar ja Niko Partanen
Digitaalisten ihmistieteiden osasto, Helsingin yliopisto

Tässä artikkelissa kokeilemme erilaisia menetelmiä kuvaavien piirteiden tuottamiseksi 151:lle alkuperäiselle Pokémonille. Tuotamme eri menetelmillä sanavektorimalleja nettikorpuksen avulla, ja luokittelemme niillä automaattisesti englannin kielen adjektiiveja sen perusteella, kuinka ominaisia ne ovat tietylle Pokémonille. Kokeidemme perusteella voimme todeta, että sanavektorimalli toimii vain, jos se on tuotettu Pokémon-korpuksen avulla. Valmiiksi tuotetut mallit eivät pysty kuvaamaan Pokémoneja hyvin. Word2Vec-malli tuottaa parempia tuloksia kuin fastText-malli. Lisäksi kokeilemme laajentaa jokaisen Pokémonin piirteiden luetteloa automaattisesti. Mikään menetelmistä ei kuitenkaan toimi riittävällä tarkkuudella, ja eri semanttisissa malleissa on kaikissa huomattava määrä ongelmia.

Avainsanat: Pokémon, sanavektorimallit, semantiikka

Johdanto

Substantiiveille tyypillisiä piirteitä sisältävien tietokantojen (*knowledge bases*) käyttö on ollut laskennallisen luovuuden tutkimuksen ytimessä jo pitkään. Piirteillä tarkoitetaan tässä yhteydessä sellaisia adjektiiveja, jotka kuvaavat tiettyä substantiivia hyvin, esimerkiksi *karvainen* on *koiraa* kuvaava piirre. Tällaiset tietokannat ovat osoittautuneet hyödyllisiksi tuotettaessa koneellisesti erityyppistä luovaa kieltä, kuten metaforia (Veale & Hao, 2007), runoja (Hämäläinen, 2018) tai arvoituksia (Ritchie, 2003).

Tässä artikkelissa esitämme uudenlaisen lähestymistavan tällaisen tietokannan rakentamiseksi automaattisesti 151:lle alkuperäiselle Pokémonille. Lähestymistapamme soveltuu tilanteisiin, joissa on käytettävissä vain rajoitettu määrä dataa. Tuloksena olevaa tietokantaa voidaan käyttää tulevaisuudessa Pokémoniin perustuvan luovan kielen, kuten vertausten ja metaforien luomiseen (esim. *Söpö kuin Pikachu* tai *hämmentynyt kuin Psyduck*).

Pokémonia on tutkittu tieteellisin menetelmin aiemminkin (Salter et al., 2019; Geissler et al., 2020; Vaterlaus et al., 2019). Aihetta ei kuitenkaan ole tutkittu aikaisemmin laajamittaisesti kieliteknologian näkökulmasta. Pokémonien nimet ovat kuitenkin yllättävän ongelmallisia nykyisille kieliteknologisille menetelmille, kuten osoitamme tässä artikkelissa.

Stereotyyppistä tietoa on onnistuttu eristämään aikaisemmin onnistuneesti datasta (Veale & Hao, 2008). Heidän menetelmänsä perustui Google-hakukoneen käyttämiseen stereotyyppisten adjektiivi-substantiivi-suhteiden kartoittamiseen käyttämällä "AS *adjektiivi* AS [a / an] *substantiivi*" -hakua. Tällainen menetelmä vaatii kuitenkin paljon dataa, jotta se toimisi, ja tällaisen kyselyn käyttäminen kohtuullisen kokoisessa korpuksessa ei omien kokemustemme perusteella juurikaan tuota tuloksia.

Erisnimille, tai tarkemmin sanottuna julkisuuden henkilöille, yksinkertaisin tapa rakentaa tällainen tietokanta on ollut manuaalinen annotaatio. Manuaalista lähestymistapaa on käytetty NOC-listan (*Non-Official Characterization*) tuottamiseen (Veale, 2016). NOC-lista sisältää tietoa erilaisista piirteistä, jotka kuvaavat julkisuuden henkilöitä. Vaikka NOC-lista on arvokas resurssi laskennallisen luovuuden mallintamisessa, me haluamme kehittää automatisoidun menetelmän samanlaisen tietokannan tuottamiseksi Pokémonille.

NOC-listan sisältämiä piirteitä on laajennettu automaattisesti datalähtöisin menetelmin (Alnajjar et al., 2017). Vaikka tämä menetelmä on askel kohti haluttua suuntaa siinä mielessä, ettei se vaadi sitä, että erisnimet, joille piirteitä laajennetaan, esiintyisivät massiivisessa korpuksessa, se nojaa edelleen louhittuihin assosiaatioihin adjektiivisten piirteiden välillä ja käsin annotoituun luetteloon julkisuuden henkilöiden piirteistä. Menetelmä ei siis kykene tuottamaan piirteitä täysin uusille, aikaisemmin NOC-listassa kuvaamattomille, henkilöille.

Ehdotamme lähestymistavaksi menetelmää Pokémonien ominaisuuksien eristämiseksi automaattisesti datasta, joka toimii hyvin pienessä korpuksessa. Lisäksi käytämme suurempaa Pokémon-korpusta luokitellaksemme automaattisesti Pokémonien piirteet niin, että korkeampi sijoitus annetaan piirteille, jotka kuvaavat parhaiten tiettyä Pokémonia.

Tuottamamme korpukset[1] ja koneoppimismallit[2] on julkaistu avoimesti Zenodossa. Tämä on tärkeää, sillä valmiiksi tuotetut sanavektorimallit eivät kykene kuvaamaan Pokémoneja kovinkaan hyvin.

Data ja esiprosessointi

Käytämme Wikidataa[3] lähteenä Pokémonien piirteille. Vaikka Wikidata ei itsessään sisällä tietoa Pokémonien piirteistä, se sisältää yksiselitteiset linkit Giantbomb-sivuston[4] Pokémonien tietosivuihin. Me käytämme Wikidatan listaa nimeltä *Pokémon introduced in*

[1] https://dx.doi.org/10.5281/zenodo.4552785
[2] https://dx.doi.org/10.5281/zenodo.4554478
[3] https://www.wikidata.org/
[4] https://www.giantbomb.com/

Generation I[5] saadaksemme tarvitsemamme Giantbomb-linkit jokaiselle 151:lle alkuperäiselle Pokémonille.

Giantbomb on verkkosivusto, jossa on tietoa videopelihahmoista. Toisin kuin esimerkiksi Bulbapedia[6], se tarjoaa lyhyen kuvauksen, joka sisältää hyödyllisiä tietoja, kuten ominaisuuksia ja fyysisiä kykyjä, kuvaamatta liian syvällisesti Pokémonin käyttöä videopeleissä. Tämä data ei kuitenkaan ole rakenteellista, vaan se on pikemminkin vapaamuotoista tekstiä. Nämä tiedot muodostavat pienen Pokémon-kuvauskorpuksemme.

Pokémonien piirteiden automaatista luokittelemista varten lataamme internetistä suuren määrän Pokémoneista kirjoitettuja tekstejä. Monet Wikipedian kaltaiset lähteet ovat liian neutraaleja kuvatakseen mitään kannaltamme mielekästä Pokémoneista, Pokédex-merkinnät ovat puolestaan yleensä liian lyhyitä ja riittämättömiä tarpeidemme kannalta. Pokémon-TV-ohjelman tekstityksien käytössä on omat ongelmansa, sillä teksti on hyvin audiovisuaalista, mikä tarkoittaa sitä, että suurin osa kuvaavasta kerronnasta tapahtuu videon, ei tekstin välityksellä. Fanfiction-sivusto[7] sisältää paljon fanien kirjoittamia tarinoita Pokémoneista, ja se on korpuksena tarpeidemme kannalta hyvä vaihtoehto, sillä tarinat sisältävät paljon Pokémoneja kuvaavia piirteitä.

Fanfiction-resurssin suurin ongelma on se, että monet tarinat on huonolla kielellä (sekä tyylillisesti että kieliopillisesti) ja että tarinoita on kirjoitettu useilla eri kielillä. Tämän ongelman ratkaisemiseksi haemme sivustolta tarinoita *pokemon*-hakusanalla rajaten hakuosumat niihin tarinoihin, jotka ovat englanninkielisiä ja joissa on vähintään 10.000 sanaa. Tällaisia fanien kirjoittamia Pokémon-tarinoita sivustolla on 8011 kappaletta. Lataamme sivustolta vain ne tarinat, jotka täyttävät nämä kriteerit. Tämä muodostaa suuremman Pokémon-tarinakorpuksemme, jonka tokenisoimme lause- ja sanatasolla käyttäen NLTK-työkalua (Bird et al., 2009).

Pokémonien piirteiden eristäminen datasta

Kokeilimme useita erilaisia menetelmiä kunkin Pokémonin piirteiden erottamiseksi. Ensiksi käytämme TF-IDF (term frequency–inverse document frequency) -menetelmää, jolla Pokémonien piirteet erotellaan ja luokitellaan niitä kuvailevasta Pokémon-kuvauskorpuksesta. Vertaamme TF-IDF-menetelmän tuloksia erilaisiin menetelmiin, jotka käyttävät sanavektoreita semanttisen läheisyyden ja samankaltaisuuden mallintamisessa. Semanttista läheisyyttä mallintaaksemme rakennamme logaritmisen todennäköisyysmatriisin (*log-likelihood*) termien välisistä suhteista, joka perustuu termien yhteisesiintymiin

[5] https://www.wikidata.org/wiki/Q3245450
[6] https://bulbapedia.bulbagarden.net/
[7] https://www.fanfiction.net/

korpuksessa käyttäen Meta4Meaning-mallia (Xiao et al., 2016). Kuvataksemme semanttista samankaltaisuutta käytämme word2vec-mallia (Mikolov et al., 2013) ja fastText-mallia (Bojanowski et al., 2016). Kokeilemme käyttää sekä valmiiksi suuresta tekstimassasta tuotettuja malleja että meidän Pokémon-tarinakorpuksestamme tuotettuja malleja, nähdäksemme kuinka suuri ero valmiiden yleisten mallien ja Pokémon-spesifien mallien välillä on.

Keräämme jokaiselle Pokémonille alustavan adjektiiveista koostuvan piirrejoukon Pokémon-kuvauskorpuksesta esikäsittelemällä korpuksen käyttämällä spaCy:ä (Honnibal & Johnson, 2015). Säilytämme jokaisen pokemonin kuvauksissa olleet adjektiivit, ja poistamme muiden sanaluokkien sanat. Tämä vaihe tuottaa listan piirteistä, joita korpuksessa käytettiin kuvaamaan kutakin Pokémonia. Se sisältää kuitenkin myös hyvin yleisiä adjektiiveja, kuten *original* (alkuperäinen), ja joissakin tapauksissa Pokémonille ei tule yhtään adjektiiveja korpuksessa olevien hyvin lyhyiden kuvausten vuoksi. Esimerkkinä *Pikachulle* kerätyt ominaisuudet sisälsivät sanoja kuten: *electric* (sähköinen), *petite* (pieni), *close* (läheinen), *cute* (söpö), *yellow* (keltainen), *high* (korkea), ... *first* (ensimmäinen), *electric* (sähköinen).

Seuraavaksi tutkimme tapoja järjestää ja laajentaa kunkin Pokémonin piirrejoukkoa. Haluamme järjestää piirteet siten, että Pokémonia parhaiten kuvaavat piirteet nousevat kärkeen ja huonoimmin kuvaavat piirteet siirtyvät piirrejoukon hännille. Ensimmäisessä menetelmässä käytämme TF-IDF-menetelmää, jossa rakennamme TF-IDF-matriisin Pokémon-kuvauskorpuksesta käsittelemällä kutakin Pokémonia dokumenttina ja niiden kuvauksia dokumenttien ominaisuuksina. Käytämme Scikit-learn-kirjastoa (Pedregosa et al., 2011) TF-IDF-matriisin tuottamiseksi. Intuitio tämän takana on, että TF-IDF pystyisi automaattisesti päättelemään kunkin piirteen merkittävyyden jokaiselle Pokémonille. Tämän seurauksena malli antaa meille luettelon jokaisesta piirteestä sekä tiedon niiden merkityksestä Pokémonille. Tämä on hyvin yksinkertainen tapa luokitella Pokémonien piirteet käyttämättä suurempaa Pokémon-tarinakorpusta. Käyttämällä TF-IDF:n palauttamia merkittävyysarvoja edellisessä vaiheessa kerättyjen ominaisuuksien luokittelussa saamme seuraavat järjestetyt ominaisuudet Pikachulle: *loveable* (rakastettava), *onomatopoetic* (onomatopoeettinen), *prolific* (tuottelias), *stubborn* (itsepäinen), *superlative* (ylivoimainen), *unbeknownst* (tietämätön), … , *15th* (viidestoista).

Seuraavissa vaiheissa luokittelemme kerätyt adjektiiviset piirteet semanttisen läheisyyden ja samankaltaisuuden kautta sanavektorimallien avulla. Kokeilemme käyttää kutakin menetelmää kahdella eri tavalla: etukäteen tuotetulla sanavektorimallilla, joka on tuotettu Wikipedian tai Common Crawlin pohjalta, sekä mallilla, jonka tuotamme Pokémon-tarinakorpuksemme pohjalta.

Noudatamme Xiaon ja kumppaneiden (2016) kuvaamaa lähestymistapaa tuottaaksemme läheisyysmatriisin. Laskemme yksinkertaisen logaritmisen todennäköisyyden kahden sanan välisen suhteen mittaamiseksi sanojen omien taajuuksien sekä korpuksessa havaittujen yhteisesiintymien taajuuksien avulla. Käytämme tähän tarkoitukseen ukWac-korpusta (Ferraresi et al., 2008) yleisenä korpuksena. Tuotamme yhden läheisyysmallin käyttämällä yleistä korpusta ja toisen mallin käyttämällä Pokémon-tarinakorpusta. Tulosten perusteella voimme sanoa, että mikään Pokémon ei kuvaudu yleisessä mallissa lukuun ottamatta kahta Pokémonia, *Persiania* ja *Dittoa*. Tämä johtuu siitä, että sekä *persian* (persialainen) että *ditto* (yllämainittu) ovat myös tavallisia englannin kielen sanoja, joten yleisen korpuksen perusteella tuotettu malli kuvaa näiden sanojen merkitystä, ei suinkaan kyseisten Pokémonien merkitystä. *Pikachun* piirteiden järjestäminen Pokémon-tarinakorpuksen avulla tuotetun mallin mukaan antaa seuraavan järjestyksen: *electric* (sähköinen), *yellow* (keltainen), *electrical* (sähkö-), *female* (naaras), *quick* (nopea), *powerful* (voimakas), ... *exclusive* (yksinomainen), *maximum* (suurin).

Käytämme word2vec- ja fastText-malleja sanojen semanttisen samankaltaisuuden mittaamiseen. Käytämme skip-gram-mallia oletusarvoisilla hyperparametreillä sekä fastTextillä että word2vec:llä. Sanavektorimenetelmämme koostuu luettelosta piirteitä (adjektiiveista), joiden samankaltaisuutta verrataan jokaisen Pokémonin vektoriin pistetulolla. Mitä enemmän piirteen vektori on samanlainen kuin Pokémonin vektori, sitä korkeammalle se sijoittuu. Valmiiksi tuotettuina word2vec- ja fastText-malleina käytämme Kutuzovin ja kumppaneiden (2017) mallia[8] sekä Mikolovin ja kumppaneiden (2018) mallia. Pokémon-tarinakorpuksen pohjalta tuotettavassa mallissamme käytämme Gensim-kirjastoa (Řehůřek & Sojka, 2010) word2vec-mallin tuottamiseen ja fastTextin virallista kirjastoa (Bojanowski et al., 2017) fastText-mallin tuottamiseen.

Samoin kuin yleisessä läheisyysmallissa, Pokémonien nimiä ei esiintynyt valmiiksi tuotetussa word2vec-mallissa. Joka tapauksessa fastText puolestaan kykenee käyttämään tietoa sanojen sisäisistä merkkijonoista koulutusvaiheessa, joten se pystyi tuottamaan semanttisia yhtäläisyyksiä Pokémonien ja adjektiivisten piirteiden välillä. *Pikachun* piirteiden järjestäminen käyttämällä valmiiksi tuotettua fastText-mallia ja Pokémon-tarinakorpuksen perusteella tuotettuja word2vec- ja fastText-malleja antaa seuraavat järjestykset:

- fastText (valmis malli): *cute* (söpö), *chuchu*, *red* (punainen), -, *evil* (paha), *yellow* (keltainen), *Japanese* (japanilainen), … , *tumultuous* (myrskyisä), *non* (ei-)
- word2vec (Pokémon-korpus): *electric* (sähköinen), *chuchu*, -, *electrical* (sähköinen), *quick* (nopea), *yellow* (keltainen), *cute* (söpö), ..., *capable* (kykenevä), *prominent* (merkittävä)

[8] http://vectors.nlpl.eu/repository/20/3.zip

- fastText (Pokémon-korpus): *electric* (sähköinen), *chuchu, electrical* (sähköinen) *cute* (söpö), *yellow* (keltainen), *close* (läheinen), … , *prolific* (tuottelias), *15th* (viidestoista)

Piirrejoukoissa on selvästi silti ongelmallisuutta, sillä mallit eivät ennusta uusia piirteitä, vaan järjestävät Pokémon-kuvauskorpuksesta poimittuja piirteitä. Tämän takia käytimme sanavektorimalleja järjestämään kaikki englannin kielen adjektiivit kullekin Pokémonille. Emme siis enää järjestä piirteitä Pokémon-kuvauskorpuksesta vaan järjestämme kaikki Oxfordin englannin sanakirjan[9] tuntemat adjektiivit kullekin Pokémonille pistetulon avulla.

Lisäksi kokeilemme olemassa olevaa menetelmää piirrejoukon laajentamiseksi kunkin menetelmän tuloksille. Piirteiden laajennus perustuu Alnajjarin ja kumppaneiden (2017) tuottamaan dataan ja algoritmiin. Menetelmä ottaa sisään luettelon piirteistä ja tuottaa laajennetun piirrelistan käyttämällä Thesaurus Rexiä (Veale & Li, 2013). Käytämme tätä menetelmää ennustamaan lisää piirteitä syöttämällä algoritmille kunkin sanavektorimallin tuottamat kymmenen parasta adjektiivia.

Tulokset

Taulukko 1 näyttää tulokset eri Pokémoneille eri menetelmillä. Taulukossa on esitetty sanavektorimallien tulokset Oxfordin sanakirjan adjektiiveille. Valmiiksi tuotettu word2vec-malli ja yleisen korpuksen pohjalta tuotettu läheisyysmalli puuttuvat taulukosta, koska ne eivät tuottaneet lainkaan tuloksia millekään Pokémonille. Lihavoiduissa soluissa on eniten adjektiiveja, jotka ovat kuvaavia Pokémonille.

Taulukosta 1 voimme nähdä, että valmiiksi tuotettu fastText-malli ei kuvaa minkään Pokémonin semantiikkaa lainkaan. Kaiken kaikkiaan Pokémon-tarinakorpuksella tuotettu fastText näyttää järjestävän hyviä ja kuvaavia adjektiiveja piirrelistassa ensimmäisten joukkoon, mutta se takeltelee selvästi sanaston ulkopuolisten adjektiivien kanssa. Sen sijaan, että malli ei palauttaisi sanaston ulkopuolisille sanoille vektoria lainkaan, se on suunniteltu palauttamaan vektorin sanan merkkitason samankaltaisuuden perusteella. Tästä syystä *Oman* ja *Omani* (omanilainen), joita ei esiintynyt Pokémon-tarinakorpuksessa, liittyvät mallin mukaan läheisesti *Omanyteen*, koska niiden merkkitason etäisyys on pieni. Vesi-Pokémoneille *swime* (huimaava, arkaainen merkitys) saa korkeat pisteet johtuen lähinnä siitä, että se on lähellä sanaa *swim* (uida).

[9] https://www.oed.com/

Pokémon	*TF-IDF*	*Pokémon fastText*	*Valmis malli fastText*	*Pokémon word2vec*	*Pokémon läheisyys*
Parasect	back, big, dark, lower, parasite	**parasitic, poisonful, sapping, crab-like, poison**	QF, Oz, EP, XL, foe	sapping, crab-like, Polish, poison, sapped	scuttled, solar, evolved, sent, male
Omanyte	full, twisted, pokemon, original, strange	fossil, Oman, Omani, fossil-like, crab-like	JV, EP, tapu, mi, zoid	**beached, fossil, crab-like,dorsal, evolved**	fossil, caught, scald, level, prehistoric
Horsea	pokemon, original, powerful	**bubble, squirtish, high-current, splashing, swime**	QF, ray, zoid, animé, peaty	bubble, beached, high-pressured, dorsal, scald	bubble, caught, evolved, level, swimming
Arcanine	pure, true, mysterious, select, majestic	lubric, whinny, mane, canine, dismounted	EP, XL, JV, pi, glew	**whinny, earth-shaking, orange-yellow, high-pressured, scald**	back, canine, large, sent, male
Abra	original, psychic	disable, Mole, Chinglish, psychic, Minimite	Oz, ex, D., EP, Ona	**hypnotic, dinged, sapping, psychic, evolved**	psychic, teleporting, side, evolved, cast
Seaking	prominent, pokemon, original	beached, high-current, swime, dorsal, hydro	QF, JV, EP, A1, zoid	**beached, tidal, high-pressured, seismic, dorsal**	released, trapped, swimming, sent, causing
Jolteon	smallest, negative, sad, shortest, startled	mane, bristled, crackled, wagging, veed	QF, EP, XL, JV, pi	whinny, supercharged, high-pressured, pi, wagging	**evolved, electric, spiky, male, female**
Magmar	fiery, pokemon, original, smaller, intense	fire-hot, knock-on, punch, seismic, scald	foe, Oz, EP, XL, zoid	five-pointed, high-pressured, seismic, scald, hydro	**punch, fiery, sent, flame, causing**
Pidgeot	beautiful, top, wide, thick, unsuspecting	bat-wing, cawing, flappish, preened, flapped	QF, EP, XL, zoid, glew	**cawing, flapped, seismic, lightning-quick, roosting**	back, flapped, evolved, flapping, landed

Taulukko 1: Järjestyksessä 5 ensimmäistä adjektiivia kullakin menetelmällä 9 satunnaiselle Pokémonille

Kaikissa tuloksissa voimme nähdä, että joidenkin Oxfordin sanakirjan adjektiivien harvinaisuus sekoittaa malleja. Parempia tuloksia voitaisiin saavuttaa, jos adjektiiviluettelo saatiin korpuksesta kattavan sanakirjan sijaan, sillä Oxfordin sanakirja sisältää myös esimerkiksi historiallisia, vanhentuneita ja murteellisia sanoja.

On hyvin hankalaa sanoa, mikä malleista toimii parhaiten, sillä ne kaikki toimivat paremmin jollekin tietyille Pokémoneille kuin muut. Voimme kuitenkin todeta, että Pokémon-tarinakorpuksen perusteella tuotetut sanavektorimallit toimivat paremmin kuin TF-IDF:n käyttäminen piirteiden poimimiseen Pokémon-kuvauskorpuksesta tai valmiiksi tuotetun mallin käyttäminen. Word2vec näyttää tuottavan täysin vääriä osumia kuin fastText.

Taulukosta 2 näemme mallin ennustamat viisi parasta piirrettä, jotka on tuotettu piirrejoukon automaattisella laajennuksella, joka perustuu kunkin menetelmän tuottamien kymmenen parhaan piirteen luetteloihin. Mikään *Beedrillin, Exeggcuten* tai *Raichun* laajennetuista piirteistä ei ollut tarpeeksi kuvaileva, jotta se olisi voitu valita parhaaksi tulokseksi. Kaiken kaikkiaan laajennetut piirteet ovat erittäin heikkoja kuvaamaan kutakin Pokémonia. Näiden tulosten perusteella emme voi suositella automaattisen piirteiden laajentamisen käyttämistä Pokémonille, koska se näyttää suosivan ihmisille tyypillisiä piirteitä. Menetelmä ei myöskään kyennyt laajentamaan kaikkien mallien piirteitä eikä yhtään valmiiksi tuotetun fastText-mallin piirettä.

Pokémon	*TF-IDF*	*Pokémon fastText*	*Valmis malli fastText*	*Pokémon word2vec*	*Pokémon läheisyys*
Drowzee	**dangerous, knowledgeable, intelligent, ruthless, twisted**	beautiful, dreamy, raw, alluring, sensuous		amusing, funny, surprising, charming, relaxed	public, versatile, despicable, specified, needed
Magnemite	light, inconspicuous, fresh, insignificant, memorable	handsome, rugged, individual		**inflexible, fixed, boring, stolid, unchanging**	soulful, grandiose, expressive, exciting, urgent
Raichu	dismayed, amazed, horrified, outraged, surprised	grandiose, funky, twisty, crazed, exciting		grandiose, funky, twisty, crazed, exciting	sturdy, potent, raw, versatile, wealthy
Beedrill	loud, dangerous, clear, deadly, slick			bitter, divisive, alive, deadly, vulgar	frustrated, disappointed, bitter, scared, shocked
Exeggcute	beautiful, creative, innovative, varied, diverse	beautiful, creative, innovative, varied, diverse		scary, inhuman, cunning, brutal, mean	
Weezing	creative, innovative, fresh, memorable, quirky	**harmful, dangerous, deadly, slick, lethal**		**harmful, dangerous, deadly, slick, lethal**	dangerous, public, specified, needed, slick
Meowth	beautiful, professional, intelligent, expressive, versatile	**crafty, clever, funny, well-meaning, treacherous**		cunning, brutal	dominant, raised, identifying, normal, known
Ninetales	**beautiful, shiny, round, passionate, merry**	crafty, clever, funny, well-meaning, treacherous		dominant, raised, identifying, normal, known	evocative, natural, fallible, alive, feminine
Arbok	scary, dangerous, funny, intense, ruthless	fluid, dangerous, detestable, unpredictable, totalitarian		**dangerous, potent, odious, slick, carcinogenic**	dominant, feminine, identifying, damp, busted

Taulukko 2: Laajennetut piirteet 9:lle satunnaiselle Pokémonille

Loppupäätelmät

Tässä artikkelissa olemme esittäneet omat lähestymistapamme Pokémon-hahmojen piirteiden eristämiseen datasta. Tulokset näyttävät lupaavilta, vaikka ne paljastavatkin sanavektorimallien semanttisen representaation ongelmat, erityisesti valmiiksi tuotetut mallit epäonnistuvat täysin Pokémonien kuvaamisessa. Merkityksellisten piirteiden automaattinen eristäminen datasta ei ole nykymenetelmillä yksinkertaista, vaan se vaatii enemmän tutkimusta tulevaisuudessa. Lähestymistapamme on kuitenkin askel poispäin täysin manuaalisesti tuotetuista tietokannoista laskennallisessa luovuudessa ja vie tutkimusta kohti täysin automaattista metodologiaa.

Tieteellinen Pokémon-seikkailumme on vasta alkanut. Tulevaisuudessa voimme tehdä erilaisia kokeita siitä, millaiset adjektiivit tuottavat parhaita tuloksia Pokémonien piirteiden eristämisessä. Myös hybridi-lähestymistapaa voitaisiin käyttää yhdistämään kunkin yksittäisen mallin vahvuudet; mitä enemmän malleja osoittaa kohti tiettyä piirrettä, sitä todennäköisemmin se on soveltuva kuvaamaan tiettyä Pokémonia.

Tutkimuksemme perusteella voimme todeta, etteivät valmiit mallit toimi Pokémoneilla ollenkaan. On selvää, ettei Pokémon ole itsessään mitenkään ilmiönä niin poikkeava, etteikö sitä voisi mallintaa sanavektoreilla. Havaitsemamme ongelma on osa laajempaa ilmiötä, johon ei kieliteknologiassa juurikaan kiinnitetä huomiota. Jos valmiit mallit, joita käytetään jatkuvasti erilaisissa kieliteknologisissa tutkimuksissa, eivät kykene kuvaamaan Pokémoneja, mitä muita ilmiöitä ne mahtavatkaan kuvata yhtä huonosti? Ylipäätään tieteenalallamme ei kiinnitetä huomiota siihen, kuinka hyvin laskennalliset mallit toimivat silloin, kun niitä sovelletaan täysin uudenlaiseen kontekstiin.

Tässä artikkelissa tuotetut sanavektorimallit voivat olla hyödyllisiä monissa erilaisissa Pokémoniin liittyvissä laskennallisen luovuuden tehtävissä. Siksi julkaisemme mallit avoimesti Zenodossa.

Lähteet

Khalid Alnajjar, Mika Hämäläinen, Hanyang Chen, and Hannu Toivonen. 2017. Expanding and weighting stereotypical properties of human characters for linguistic creativity. In ICCC, sivut 25–32.

Steven Bird, Ewan Klein, and Edward Loper. 2009. Natural Language Processing with Python, 1st edition. O'Reilly Media, Inc.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. arXiv preprint arXiv:1607. 04606.

Piotr Bojanowski, Edouard Grave, Armand Joulin, andTomas Mikolov. 2017. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5:135–146.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukwac, a very large web derived corpus of English. Teoksessa proceedings of the 4th Web as Corpus Workshop(WAC4) Can we beat Google, sivut 47–54

Dominique Geissler, Elisa Nguyen, Daphne Theodorakopoulos, and Lorenzo Gatti. 2020. Pokérator - unveil your inner Pokémon. Teoksessa Proceedings of the Eleventh International Conference on Computational Creativity

Mika Hämäläinen. 2018. Harnessing NLG to create Finnish poetry automatically. In International Conference on Computational Creativity, sivut 9–15. Association for Computational Creativity (ACC).

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. Teoksessa proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, sivut 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Andrei Kutuzov, Murhaf Fares, Stephan Oepen, and Erik Velldal. 2017. Word vectors, reuse, and replicability: Towards a community repository of large text resources. Teoksessa proceedings of the 58th Conference on Simulation and Modelling, sivut 271–276.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301. 3781

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances Teoksessa pretraining distributed word representations. Teoksessa proceedings of the International Conference on Language Resources and Evaluation (LREC2018).

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M.

Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning Teoksessa python. Journal of Machine Learning Research, 12:2825–2830

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. Teoksessa proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, sivut 45–50, Valletta, Malta. ELRA.

Graeme Ritchie. 2003. The JAPE riddle generator: technical specification. Institute for Communicating and Collaborative Systems

Anastasia Salter, Mel Stanfill, and Anne Sullivan. 2019. But Does Pikachu Love You? Reproductive Labor in Casual and Hardcore Games. Teoksessa Proceedings of the 14th International Conference on the Foundations of Digital Games. Association for Computing Machinery, New York, NY, USA.

J Mitchell Vaterlaus, Kala Frantz, and Tracey Robecker. 2019. "Reliving my childhood dream of being a Pokémon trainer": An exploratory study of college student uses and gratifications related to Pokémon GO. International Journal of Human–Computer Interaction, 35(7):596–604

Tony Veale. 2016. Round up the usual suspects: Knowledgebased metaphor generation. Teoksessa proceedings of the Fourth Workshop on Metaphor in NLP, sivut 34–41

Tony Veale and Yanfen Hao. 2007. Comprehending and generating apt metaphors: a web-driven, case-based approach to figurative language. In AAAI, volume 2007, sivut 1471–1476

Tony Veale and Yanfen Hao. 2008. Enriching wordnet with folk knowledge and stereotypes. Teoksessa proceedings of the 4th Global WordNet Conference, Szeged, Hungary

Tony Veale and Guofu Li. 2013. Creating similarity: Lateral thinking for vertical similarity judgments. Teoksessa proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1:Long Papers), sivut 660–670.

Ping Xiao, Khalid Alnajjar, Mark Granroth-Wilding, Kat Agres, and Hannu Toivonen. 2016. Meta4meaning: Automatic metaphor interpretation using corpus-derived word associations. Teoksessa proceedings of the 7th International Conference on Computational Creativity (ICCC). Paris, France