Prepoznavanje idiomatskih besednih zvez z uporabo besednih vložitev

Tadej Škvorc[1, 2], Marko Robnik Šikonja[1]
[1]Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Večna pot 113, 1000 Ljubljana.
[2]Institut Jožef Stefan, Jamova cesta 39, 1000 Ljubljana.
tadej.skvorc@fri.uni-lj.si, marko.robnik@fri.uni-lj.si

Izvleček

Prisotnost idiomov v besedilu povzroča probleme številnim pristopom na področju obdelave naravnega jezika, saj jih računalniki težko prepoznajo. Strojno prepoznavanje takšnih izrazov še ni rešen problem. V zadnjih letih so razvili številne metode, ki lahko prepoznajo različne pomene besed glede na njihovo okolico in na podlagi tega zgradijo kontekstne vektorske vložitve besed. Takšne vložitve bi morale biti primerne za zaznavanje idiomov. Trenutni pristopi ali ne uporabljajo vektorskih vložitev ali pa uporabljajo ne-kontekstne vložitve. V delu pokažemo, kako lahko uporabimo kontekstne vložitve besed za ločevanje med dobesedno in idiomatsko rabo besed. Pokažemo, da lahko z različnimi značilkami (npr., s kontekstualnimi vektorji in razdaljami do srednjih kontekstualnih vektorjev za vsako besedo) zaznamo idiome prisotne v korpusu angleških besedil GloWbE.

Ključne besede: Večbesedni izrazi, obdelava naravnega jezika, besedilno rudarjenje, vektorske vložitve besed.

Abstract

The presence of idioms presents problems for many tasks in natural language processing, as they can be hard for computers to detect. Detecting such expressions and correctly determining their meanings has not yet been fully solved. In recent years, several methods for constructing contextual word embeddings have been proposed, which are capable of detecting the different meanings of the same word based on context. Such embeddings should be well suited to detecting idioms. Current approaches either do not use embeddings or use non-contextual embeddings. We have demonstrated that we can use contextual word embeddings to differentiate between literal and idiomatic word use. We have extracted various features (e.g. the contextual vectors and distance to the mean contextual vector for each word) and shown that they can be useful for detecting idiomatic word expressions present in the GloWbE corpus of English texts.

Keywords: Multi-word expression, natural language processing, text mining, word embeddings

1 UVOD

Prisotnost idiomov in ostalih večbesednih zvez oteži mnoge naloge obdelave naravnega jezika, kot so strojno prevajanje, analiza sentimenta in samodejno povzemanje, saj je lahko prenesen pomen besed različen od dobesednega (npr. "vreči puško v koruzo", ki pomeni obupati ali odnehati).

Idiome lahko zaznamo z uporabo slovarjev, vendar te v strojni obliki niso na voljo v vseh jezikih in niso primerni za pristope, ki morajo delovati na več jezikih hkrati. Nekateri trenutni pristopi temeljijo na uporabi modelov strojnega učenja. Savary et al. [Savary et al., 2017] prestavijo pregled trenutnih pristopov in njihove rezultate na raziskovalem izzivu PARSEME o samodejnem zaznavanju glagolskih večbesednih zvez. Takšni pristopi lahko poleg idiomov prepoznajo tudi druge večbesedne zveze. Na izzivu je najbolje deloval sistem TRANSITION [Constant and Nivre, 2016], ki je dosegel najboljše rezultate na številu različnih jezikov. Pozneje so Stodden et al. [Stodden et al., 2018] pristop izboljšali tako, da so model strojnega učenja zamenjali s konvolucijsko nevronsko mrežo. Dobre rezultate je dosegel tudi sistem MUMULS [Klyueva et al., 2017].

Tabela 1: **Izluščeni idiomi in število pojavitev v korpusu.**

Idiom	Število izluščenih povedi	Število vseh povedi
Take place	1999	47522
Under the weather	200	472
The last straw	1000	1108
Bent out of shape	200	381
Hang in there	1000	2082
Break the ice	500	658
Live and learn	200	472

Zgoraj navedeni pristopi temeljijo na strojnem učenju in potrebujejo velike korpuse besedil z ročno označenimi idiomi, ki v veliko jezikih niso na voljo. Mnogim opravilom bi bil v pomoč pristop, ki se ne zanaša na zunanje vire za zaznavanje idiomov. To lahko dosežemo z uporabo vektorskih vložitev besed. Namesto, da besede ponazorimo z znaki jih lahko ponazorimo z vektorji na takšen način, da ti vektorji odražajo pomene besed in sorodnosti med njimi. Takšne vložitve lahko pridobimo z nenadzorovanim strojnim učenjem, ki ne potrebuje ročno-označenih besedil. Za izgradnjo vektorskih vložitev obstajajo mnoge metode

[Mikolov et al., 2013, Pennington et al., 2014]. Številni avtorji so pokazali, da lahko tako pridobljeni vektorji kodirajo pomenske podobnosti in sorodnosti med besedami [Bojanowski et al., 2017, Mikolov et al., 2013]. Poleg tega lahko na podlagi pridobljenih vektorjev odgovorimo na vprašanja tipa »*A* je proti *B*, kot je *C* proti«, če poznamo vektorje besed *A*, *B* in *C*.

Vektorske vložitve lahko uporabimo za zaznavanje idiomov. Ker se besede v idiomski rabi pojavijo v drugačnih kontekstih kot v dobesedni rabi, lahko na podlagi vektorjev besed v kontekstu prepoznamo ali gre za idiomsko ali za dobesedno rabo. Gharbieh et al. [Gharbieh et al., 2016] predstavijo metodo za zaznavanje idiomov oblike glagol-samostalnik na podlagi povprečja vektorjev v okolici dvoumnih besed. Peng in Feldman [Peng and Feldman, 2015] predstavita dva podobna pristopa. Prvi temelji na skalarnem produktu vektorjev besed v kontekstu, drugi pa na njihovih kovariančnih matrikah. Oba pristopa za izgradnjo vektorjev uporabita metodo word2vec [Mikolov et al., 2013].

Pred kratkim so začeli razvijati kontekstne vektorske vložitve [Peters et al., 2018], s katerimi želimo upoštevati več možnih pomenov posamezne besede. Namesto, da vsaki besedi priredimo le en vektor, priredimo vektor vsaki pojavitvi besede, pri čimer so vektorji odvisni od okolice pojavitve besede. Glavas et al. [Glavas et al., 2019] pokažejo, da lahko z uporabo takšnih vektorjev izboljšamo delovanje številnih pristopov na področju obdelave naravnega jezika.

V našem delu predstavimo, kako lahko kontekstne vložitve besed uporabimo za zaznavanje idiomov. Pri kontekstnih vložitvah za zaznavanje idiomov ne potrebujemo vektorjev besed konteksta, saj je informacija o kontekstu vsebovana v vektorju vsake pojavitve besede. Za razliko od obstoječih metod, ki delujejo na ne- kontekstnih vektorskih vložitvah, ne potrebujemo dodatnih metod, s katerimi iz vektorjev konteksta izluščimo koristne informacije. Namesto tega lahko neposredno uporabimo vektorje posameznih besed za zaznavanje idiomov. V našem delu uporabimo ELMo kontekstne vložitve [Peters et al., 2018], naučene na korpusu "1 Billion Words Benchmark" [Chelba et al., 2013], s katerimi izračunamo vložitve za besede iz sedmih različnih idiomov, ki so prisotni v korpusu "Global Web-based English Corpus" (GloWbE) [Davies and Fuchs, 2015]. Z vizualizacijo pokažemo, da so pridobljene vložitve zmožne razlikovati med prenesenimi in dobesednimi pomeni besed. Poleg tega pokažemo, da lahko z uporabo vložitev dosežemo visoko točnost pri strojnem zaznavanju idiomov. Glavna novost našega pristopa je, da ne potrebuje velikih, ročno-označenih korpusov, ki so potrebni za ostale podobne pristope.

2 ZAZNAVANJE IDIOMOV S KONTEKSTNIMI VLOŽITVAMI BESED

Za analizo zaznavanja idiomov potrebujemo veliko podatkovno množico z označenimi idiomi. Obstoječe množice (npr. množica za raziskovalni izziv PARSEME) so majhne in vsebujejo majhno število idiomov. Problem smo rešili tako, da smo zgradili lastno podatkovno množico iz besedil prisotnih v korpusu

"Global Web-based English Corpus (GloWbE)", ki vsebuje 1.9 milijard besed iz spletnih besedil. Iz besedil smo izluščili povedi, v katerih se je pojavil eden izmed sedmih različnih angleških idiomov. Idiomi, število izluščenih povedi in število vseh povedi ki vsebujejo idiom, so prikazani v Tabeli 1.

V podatkovno množico smo dodatno vključili povedi z dobesednimi pomeni besed v idiomih. Iz korpusa smo naključno izbrali toliko povedi, da smo dobili uravnoteženo podatkovno množico (53,7% vsebuje idiome).

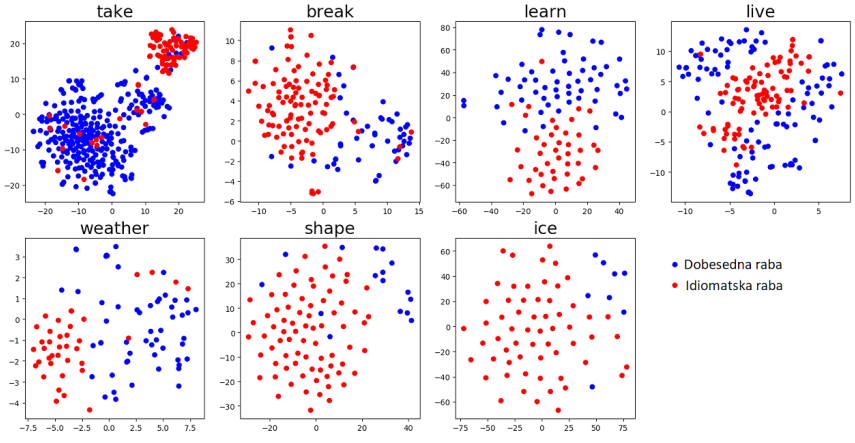
Tabela 2: **Srednje vrednosti in standardni odkloni kosinusnih razdalj do srednje vrednosti vektorjev besede.**

Beseda	Razdalja do srednje vrednosti (besedna raba)	Razdalja do srednje vrednosti (idiom)
Take	0.807 (0.070)	0.722 (0.073)
Weather	0.765 (0.055)	0.731 (0.035)
Straw	0.770 (0.010)	0.785 (0.029)
Bent	0.562 (0.011)	0.671 (0.039)
Shape	0.699 (0.126)	0.762 (0.032)
Hang	0.838 (0.055)	0.887 (0.023)
Ice	0.795 (0.014)	0.819 (0.033)
Live	0.769 (0.093)	0.845 (0.040)
Learn	0.877 (0.024)	0.839 (0.038)

Za izračun vložitev ELMo smo uporabili nevronsko mrežo, ki je bila vnaprej naučena1 na problemu jezikovnega modeliranja na korpusu "1 Billion Word Benchmark dataset [Chelba et al., 2013]". Izračunali smo vektorje dimenzije 1024, kjer smo uporabili celotno poved kot kontekst besede. Vektorje smo vizualizirali z metodo t-SNE [Maaten and Hinton, 2008]. Poleg tega smo na podlagi vektorjev izvedli klasifikacijo med idiomatskimi in dobesednimi rabami besed. Preverili smo tudi, ali je razdalja do srednjega vektorja besede koristna za zaznavanje idiomov. Rezultate predstavimo v razdelku 3.

3 REZULTATI

Za vizualizacijo smo izračunali vektorje devetih različnih besed, ki so se pojavile tako v idiomih kot dobesedno. Vektorje smo pretvorili v dve dimenziji z metodo t-SNE. S tem lahko vidimo, kako se vektorji idiomov razlikujejo od vektorjev besed v njihovi dobesedni rabi. Pri analizi vidimo, da lahko z vektorji ELMo v nekaterih primerih ločim med idiomi in dobesedno rabo besed. Na Sliki 1 so vektorji besede "take" v idiomu "take place" jasno ločeni od ostalih rab besede. Vektorji, ki se prekrivajo verjetno predstavljajo ostale prenesene pomene besede "take". Podobno velja za ostale besede, čeprav ločitev ni tako jasno razvidna.



Slika 1: **Vizualizacija vektorjev besede "take" , "break", "learn", "live", "weather", "shape" in "ice"**

Prav tako smo preverili, ali je razdalja posameznega vektorja do srednje vrednosti vseh vektorjev dovolj za zaznavanje idiomov. Statistično so dobesedne rabe besed bolj pogoste kot idiomske, torej bi morale biti bližje srednji vrednosti vseh vektorjev. Ta lastnost ne velja za našo podatkovno množico, saj smo jo zgradili tako, da je uravnotežena. Zaradi tega smo za ta poskus srednjo vrednost vektorjev izračunali iz novega naključnega vzorca vseh besed v korpusu. Nato smo za 100 dobesednih in 100 idiomatskih rab besed iz naše podatkovne množice izračunali kosinusno razdaljo do srednjega vektorja. Povprečene razdalje in standardni odkloni so prikazani v Tabeli 2.

Tabela 3: **Rezultati klasifikacije z 10-kratnim prečnim preverjanjem.**

Klasifikacijski model	Povprečna klasifikacijska točnost	95 % interval zaupanja
Naključni gozdovi	0.928	+/- 0.045
Metoda podpornih vektorjev	0.926	+/- 0.060
Extreme gradient boosting	0.940	+/- 0.049

Iz tabele je razvidno, da se razdalje razlikujejo glede na rabo besede. Pri nekaterih besedah se izkaže, da so dobesedne rabe besed bližje srednji vrednosti vektorjev kot idiomske rabe. To nasprotuje naši hipotezi, vendar prisotnost razlik nakazuje da bi razdalje še vedno lahko bile koristne pri zaznavanju idiomov. Ena možna razlaga je, da so v nekaterih primerih srednje vrednosti obeh skupin blizu druga drugi, v katerem primeru bo razdalja odvisna predvsem od standardnega odklona distribucij vektorjev obeh skupin. Če so vektorji idiomatskih rab skoncentrirani skupaj, vektorji dobesednih rab pa daleč narazen bodo idiomatske rabe bližje srednji vrednosti vseh vektorjev.

Preverili smo tudi, ali lahko uporabimo klasifikacijske modele na vektorskih vložitvah za zaznavanje idiomov. Uporabili smo sledeče modele: naključni gozdovi [Breiman, 2001], metoda podpornih vektorjev [Suykens and Vandewalle, 1999] in extreme gradient boosting [Chen et al., 2015]. Rezultati 10-kratnega prečnega preverjanja so prikazani v Tabeli 3.

Najboljše rezultate smo dosegli z metodo extreme gradient boosting. Z vsemi metodami smo dosegli visoke rezultate. Čeprav je naša podatkovna množica je preprosta, zaradi česar bi bili rezultati na zahtevnejših primerih verjetno slabši, to nakazuje da so kontekstne vektorske vložitve primerne za zaznavanje idiomov.

4 ZAKLJUČEK

Zaznavanje idiomov je pomembna za veliko opravil na področju obdelave naravnega jezika. Pokazali smo, da lahko kontekstne vektorske vložitve uporabimo za zaznavanje idiomov, pri čemer ne potrebujemo ročno označenih korpusov. Na umetni podatkovni množici z vektorji lahko ustrezno ločimo dobesedne in idiomske rabe in s klasifikacijo dosežemo 94 % točnost.

V nadaljnem delu nameravamo razširiti evaluacijo z uporabo večje podatkovne množice in preveriti, ali lahko z vektorskimi vložitvami izboljšamo delovanje trenutnih pristopov. Trenutna evaluacija je omejena na le sedem idiomov, zaradi česar je težko vedeti, kako dobro naš pristop deluje na vseh idiomih, predvsem na takšnih, ki se v besedilih redko pojavijo.

ZAHVALA

Raziskovalno delo je bilo sofinancirano s strani Javne agencije za raziskovalno dejavnost Republike Slovenije, št. projekta P2-0209 in P2-0103.

LITERATURA

[1] [Bojanowski et al., 2017] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

[2] [Breiman, 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[3] [Chelba et al., 2013] Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., and Robinson, T. (2013). One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

[4] [Chen et al., 2015] Chen, T., He, T., Benesty, M., Khotilovich, V., and Tang, Y. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, pages 1–4.

[5] [Constant and Nivre, 2016] Constant, M. and Nivre, J. (2016). A transition-based system for joint lexical and syntactic analysis. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 161–171.

[1] https://tfhub.dev/google/elmo/2

[6] [Davies and Fuchs, 2015] Davies, M. and Fuchs, R. (2015). Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide*, 36(1):1–28.

[7] [Gharbieh et al., 2016] Gharbieh, W., Bhavsar, V., and Cook, P. (2016). A word embedding approach to identifying verb-noun idiomatic combinations. In *Proceedings of the 12th Workshop on Multiword Expressions*, pages 112–118, Berlin, Germany. Association for Computational Linguistics.

[8] [Glavas et al., 2019] Glavas, G., Litschko, R., Ruder, S., and Vulic, I. (2019). How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. *arXiv preprint arXiv:1902.00508*.

[9] [Klyueva et al., 2017] Klyueva, N., Doucet, A., and Straka, M. (2017). Neural networks for multi-word expression detection. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 60–65.

[10] [Maaten and Hinton, 2008] Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

[11] [Mikolov et al., 2013] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[12] [Peng and Feldman, 2015] Peng, J. and Feldman, A. (2015). Automatic idiom recognition with word embeddings. In *Information Management and Big Data*, pages 17–29. Springer.

[13] [Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

[14] [Peters et al., 2018] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proc. of NAACL*.

[15] [Savary et al., 2017] Savary, A., Ramisch, C., Cordeiro, S., Sangati, F., Vincze, V., QasemiZadeh, B., Candito, M., Cap, F., Giouli, V., and Stoyanova, I. (2017). The parseme shared task on automatic identification of verbal multiword expressions.

[16] [Stodden et al., 2018] Stodden, R., QasemiZadeh, B., and Kallmeyer, L. (2018). Trapacc and trapaccs at parseme shared task 2018: Neural transition tagging of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 268–274.

[17] [Suykens and Vandewalle, 1999] Suykens, J. A. and Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300.

▪

*T*adej Škvorc je mladi raziskovalec in doktorski študent na Fakulteti za računalništvo in informatiko Univerze v Ljubljani. Ukvarja se s področjem obdelave naravnega jezika in umetne inteligence.

▪

Marko Robnik-Šikonja je redni profesor in predstojnik Katedre za umetno inteligenco Fakulteti za računalništvo in informatiko Univerze v Ljubljani. Raziskovalno se ukvarja s področji umetne inteligence, strojnega učenja, obdelave naravnega jezika in analize omrežij. Je avtor več kot 100 znanstvenih publikacij, ki so bile citirane več kot 4000-krat.