LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

PROGRAMMING LANGUAGES AND ARTIFICIAL INTELLIGENCE

Titel der Arbeit

Ruben Triwari

Bachelorarbeit
im Studiengang ’Informatik plus Mathematik’

Betreuer: Prof. Dr. Johannes Kinder

Mentor: Moritz Dannehl, M.Sc.

Ablieferungstermin: 28. August 2024

Inhaltsverzeichnis

Abstract

1 Einführung

In den letzten Jahren gab es große Fortschritte in der natürlichen Sprachverarbeitung, besonders hervorzuheben sind Large Language Models die sich mittlerweile in vielen Bereichen der Informatik in die Lösungsansätze für Problemen in jeweiligen Bereichen eingeschlichen haben. Diese Arbeit untersucht nun, ob diese Fortschritte in der natürlichen Sprachverarbeitung eine Hilfestellung leisten können um Source Code Funktionen semantisch sinvoll in einen Vektor mit reelwertigen Zahlen zu codieren. Diese Vektoren können dann später als Label verwendet werden um ein Modell zu trainieren was Binary Code als Input nimmt und diesen ebenfalls in einen semantischen Vektor mit reelwertigen Zahlen codiert. Das resultierende Modell kann hinterher verwendet werden um Reverse Engeneering zu erleichtern. Ein einfaches Beispiel ist folgendes: Man stelle sich vor, dass man eine Funktion die in Binary Code vorliegt, mühsehlig manuell verstanden was für eine Aufgabe die Funktion in der Code Base hat. Nun kann man diese Funktion codieren und über die Gesamte Code Base einen Nearest Neighbor Search durchführen und all ähnlichen Funktionen ausgeben lassen. Das spart zeit, denn nun hat man eine Idee was diese anderen Funktionen für eine Aufgabe in der Code Base erfüllen könnten.
Das oben berschriebene Problem Source Code Vektoren in sinvoll semantische reelwertige Vektoren zu codieren ist sehr ähnlich zu einen Problem in der natürlichen Sprachverarbeitung und dort bereits gelöst. Die rede ist von dem Problem einen gegebenen Satz in einen semantisch sinvollen Vektor abzubilden. Es ist nageliegend zu versuchen dieses Ergebnis der natürlichen Sprachverarbeitung zu benuzten um eine Lösung für unser Problem zu konstruieren. Die intuivste Idee ist es einfach die Funktionsnamen, die in natürlicher Sprache verfasst sind als beschreibung der Funktion zu verwenden. Diese Beschreibunf können wir nun mühelos codieren, da sie in natürlicher Sprache vorliegt. Eine zweite Idee ist, die Kommentare der Funktionen, die in natürlicher Sprache verfasst sind, als Beschreibung der Funktion zu verwenden. Am viel versprechsten ist es die Funktionen von einen Large Language Modell in natürlicher Sprache beschreiben zu lassen. Als letztes habe ich noch ein bestehendes Modell Code2Vec verwendet und es für dieses Problem angepasst.

2 Grundlagen und Termini

2.1 Maschienelles Lernen

In diesen Abschnitt wird zunächst maschinelles lernen definiert und dann darauf aufbauend grundlegende Trainingsarten vorgestellt. Heutzutage ist maschienelles Lernen weitverbreitet und wird nahezu in jeden Bereich der Informatik verwendet. Maschinelles Lernen wird überall eingesetzt wo eine anayltische Lösung eines Problems zu aufwendig oder gar überhauptnicht existiert.

2.1.1 Definition

Diese Lösung durch maschinelles Lernen versucht aus den Daten ein Muster abzuleiten. Bei einer endlichen Menge an Daten ist klar, dass das resultierende Modell nur eine approximation der gesuchten Lösung ist. Trotz des bekanntheitsgrades, gibt es den irreglauben, dass maschinelles lernen nur was mit Neuronalen Netzwerken zu tun hat, diese Anahme ist im allgemeinen falsch. Im folgenden wird eine allgeimeine Defintion von maschinellen Lernen vorgestellt:

Definition 1 *Sei X eine beliebige Input Menge, Y eine beliebige Output Menge, $f \in \{X \to Y\}$ die gesuchte Lösung des Problems, $\mathbb{D}$ eine beliebige Menge aus gegebenen Datenpunkten,*
$H_1 \subset \{X \to Y\}$ ein Hypothesenraum, und $A_1 : \mathscr{P}(\{X \to Y\}) \times \mathscr{P}(\mathbb{D}) \to \{X \to Y\}$ ein Lernalgorithmus. Dann ist das ziel, bei gegebenen Daten, den Hypothesenraum H_1 und den Lernalgorithmus A_1 so zu wählen, sodass

$$A_1(H_1, \mathbb{D}) \approx f.$$

Maschinelles lernen ist also die suche nach einen Lernalgorithmus und Hypothesenraum, die dann in kombination mit gegebenen Daten, die optimale Lösung approximieren. Dabei ist hervorzuheben das der Datensatz das Herzstück jeder Problemstellung im Bereich des maschinellen Lernens ist. Ist der Datensatz zu klein oder überhaupt nicht representativ für das gegebene Probelm, wird der Lernalgorithmus die falschen Muster erkennen und dadurch eine fehlerhafte approximation produzieren.

2.2 Sentence Transformer

2.3 Large Language Models

2.4 Code2Vec

3 Methodik

3.1 Datensatz

Im Maschinellen lernen hat der Datensatz bzw. die Trainingsdaten den größten Einfluss auf die güte des Modells.

3.2 Datenpipeline

3.3 Stabilität von SentenceTransformer

- Stabilität von Sentence Transformer

- Kommentare von Funktionen um Embeddings zu generieren
- Funktionsnamen von Funktionen um Embeddings zu generieren
- Code2Vec um Embeddings zu generieren
- CodeLlama Erklärungen von Funktionen um Embeddings zu generieren
- Evaluierung durch tSNE-Plots
- Evaluierung durch Experten
- Evaluierung durch Formel

$I_k : \mathbf{N} \times \mathbf{N} \times \mathbf{N}^k \to [0,1]$

$$I_k(x,i,v) = \begin{cases} 1 & , \exists j \in \mathbf{N} : x = v_j \wedge i = j \\ \frac{1}{2} & , \exists j \in \mathbf{N} : x = v_j \wedge i \neq j \\ 0 & , \text{otherwise} \end{cases}$$

$E_k : \mathbf{N}^k \times \mathbf{N}^k \to [0,1]$

$$E_k(u,v) = \frac{1}{G_k} \sum_{i=1}^{k} \frac{I_k(u_i,i,v_i)}{log_2(i+1)}$$

wo $G_k := \sum_{i=1}^{k} \frac{1}{log_2(i+1)}$.

$CMP_k : \mathbf{R}^{N \times l} \times \mathbf{R}^{N \times l} \times \{\mathbf{R}^l \times \mathbf{R}^{N \times l} \to \mathbf{N}^k\} \times \{\mathscr{P}([0,1]) \to [0,1]\} \to [0,1]$

$$CMP_k(X,Y,f_k,agg) = agg(\{E_k(f_k(X_{i,j},X), f_k(Y_{i,j},Y)) | j \in \{1,2,3,\ldots N\}\})$$

13 Conclusion

14 Notes on form

14.1 Formatting

This LaTeX template uses the following formatting:

- font: Linux Libertine O (alternatively: Times New Roman)
- font size: 12 pt
- left and right margin: 3.5 cm, top and bottom margin: 3 cm
- align: left
- line spacing: one and a half (alternative: 15 pt line spacing with 12 pt font size)

When implementing the specifications in Word, it is essential to define style sheets.

14.2 Citation

The citation method follows the author-year system. Place reference is in the text, footnotes should only be used for explanations and comments. The following notes are taken from the *language* bibliography template from `ron.artstein.org`:

The *Language* style sheet makes a distinction between two kinds of in-text citations: citing a work and citing an author.

- Citing a work:
  - Two authors are joined by an ampersand (&).
  - More than two authors are abbreviated with *et al.*
  - No parentheses are placed around the year (though parentheses may contain the whole citation).
- Citing an author:
  - Two authors are joined by *and*.
  - More than two authors are abbreviated with *and colleagues*.
  - The year is surrounded by parentheses (with page numbers, if present).

To provide for both kinds of citations, `language.bst` capitalizes on the fact that `natbib` citation commands come in two flavors. In a typical style compatible with `natbib`, ordinary commands such as `\citet` and `\citep` produce short citations abbreviated with *et al.*, whereas starred commands such as `\citet*` and `\citep*` produce a citation with a full author list. Since *Language* does not require citations with full authors, the style `language.bst` repurposes the starred commands to be used for citing the author. The following table shows how the `natbib` citation commands work with `language.bst`.

Command	Two authors	More than two authors
`\citet`	Hale & White Eagle (1980)	Sprouse et al. (2011)
`\citet*`	Hale und White Eagle (1980)	Sprouse and colleagues (2011)
`\citep`	(Hale & White Eagle 1980)	(Sprouse et al. 2011)
`\citep*`	(Hale und White Eagle 1980)	(Sprouse and colleagues 2011)
`\citealt`	Hale & White Eagle 1980	Sprouse et al. 2011
`\citealt*`	Hale und White Eagle 1980	Sprouse and colleagues 2011
`\citealp`	Hale & White Eagle 1980	Sprouse et al. 2011
`\citealp*`	Hale und White Eagle 1980	Sprouse and colleagues 2011
`\citeauthor`	Hale & White Eagle	Sprouse et al.
`\citeauthor*`	Hale und White Eagle	Sprouse and colleagues
`\citefullauthor`	Hale und White Eagle	Sprouse and colleagues

Authors of *Language* articles would typically use `\citet*`, `\citep`, `\citealt` and `\citeauthor*`, though they could use any of the above commands. There is no command for giving a full list of authors.

Bibliography

The bibliography of this template includes the references of the *language* stylesheet as a sample bibliography.

15 General Addenda

If there are several additions you want to add, but they do not fit into the thesis itself, they belong here.

15.1 Detailed Addition

Even sections are possible, but usually only used for several elements in, e.g. tables, images, etc.

16 Figures

16.1 Example 1

16.2 Example 2

Abbildungsverzeichnis

Tabellenverzeichnis

Literatur

BUTT, MIRIAM, und WILHELM GEUDER (Hg.) 1998. *The projection of arguments: Lexical and compositional factors*. Stanford, CA: CSLI Publications.

CROFT, WILLIAM. 1998. Event structure in argument linking. In Butt & Geuder, 21–63.

DONOHUE, MARK. 2009. Geography is more robust than linguistics. Science e-letter, 13 August 2009. URL `http://www.sciencemag.org/cgi/eletters/324/5926/464-c`.

DORIAN, NANCY C. (Hg.) 1989. *Investigating obsolescence*. Cambridge: Cambridge University Press.

GROPEN, JESS; STEVEN PINKER; MICHELLE HOLLANDER; RICHARD GOLDBERG; und RONALD WILSON. 1989. The learnability and acquisition of the dative alternation in English. *Language* 65.203–57.

HALE, KENNETH, und JOSIE WHITE EAGLE. 1980. A preliminary metrical account of Winnebago accent. *International Journal of American Linguistics* 46.117–32.

HASPELMATH, MARTIN. 1993. *A grammar of lezgian*. Walter de Gruyter.

HYMES, DELL H. 1974a. *Foundations in sociolinguistics: An ethnographic approach*. Philadelphia: University of Pennsylvania Press.

HYMES, DELL H. (Hg.) 1974b. *Studies in the history of linguistics: Traditions and paradigms*. Bloomington: Indiana University Press.

HYMES, DELL H. 1980. *Language in education: Ethnolinguistic essays*. Washington, DC: Center for Applied Linguistics.

MINER, KENNETH. 1990. Winnebago accent: The rest of the data. Lawrence: University of Kansas, MS.

MORGAN, TJH; NT UOMINI; LE RENDELL; L CHOUINARD-THULY; SE STREET; HM LEWIS; CP CROSS; C EVANS; R KEARNEY; I DE LA TORRE; ET AL. 2015. Experimental evidence for the co-evolution of hominin tool-making teaching and language. *Nature communications* 6.

PERLMUTTER, DAVID M. 1978. Impersonal passives and the unaccusative hypothesis. *Berkeley Linguistics Society* 4.157–89.

POSER, WILLIAM. 1984. *The phonetics and phonology of tone and intonation in Japanese*. Cambridge, MA: MIT Dissertation.

PRINCE, ELLEN. 1991. Relative clauses, resumptive pronouns, and kind-sentences. Paper presented at the annual meeting of the Linguistic Society of America, Chicago.

RICE, KEREN. 1989. *A grammar of Slave*. Berlin: Mouton de Gruyter.

SALTZMAN, ELLIOT; HOSUNG NAM; JELENA KRIVOKAPIC; und LOUIS GOLDSTEIN. 2008. A task-dynamic toolkit for modeling the effects of prosodic structure on articulation. *Proceedings of the 4th International Conference on Speech Prosody (Speech Prosody 2008)*, Campinas, 175–84. URL `http://aune.lpl.univ-aix.fr/~sprosig/sp2008/papers/3inv.pdf`.

VAN DER SANDT, ROB A. 1992. Presupposition projection as anaphora resolution. *Journal of Semantics* 9.333–77.

SINGLER, JOHN VICTOR. 1992. Review of Melanesian English and the Oceanic substrate, by Roger M. Keesing. *Language* 68.176–82.

SPROUSE, JON; MATT WAGERS; und COLIN PHILLIPS. 2011. A test of the relation between working memory capacity and syntactic island effects. *Language*, to appear.

STOCKWELL, ROBERT P. 1993. Obituary of Dwight L. Bolinger. *Language* 69.99–112.

SUNDELL, TIMOTHY R. 2009. Metalinguistic disagreement. Ann Arbor: University of Michigan, MS. URL `http://faculty.wcas.northwestern.edu/~trs341/papers.html`.

TIERSMA, PETER M. 1993. Linguistic issues in the law. *Language* 69.113–37.

WILSON, DEIRDRE. 1975. *Presuppositions and non-truth-conditional semantics*. London: Academic Press.

YIP, MOIRA. 1991. Coronals, consonant clusters, and the coda condition. *The special status of coronals: Internal and external evidence*, hrsg. von Carole Paradis und Jean-François Prunet, 61–78. San Diego, CA: Academic Press.