Kampweg 55
3769 DE Soesterberg
Postbus 23
3769 ZG Soesterberg

www.tno.nl

T +31 88 866 15 00
F +31 34 635 39 77

TNO-rapport

Update rapport: Natural Language Processing technieken voor het opstellen en verbeteren van mappings

Datum	28 november 2022
Auteur(s)	Maaike de Boer Roos Bakker Linda Oosterheert Maaike Burghoorn Quirine Smit
Aantal pagina's	5 (incl. bijlagen)
Projectnaam	APPL.AI Skills Matching 2022
Projectnummer	060.51127

1 Natural Language Processing technieken voor het opstellen en verbeteren van mappings

In het Skills Matching project zijn in WP1 van 2021 verschillende beroepsclassificaties gekoppeld. Deze koppelingen zijn tot stand gekomen door handmatige mappings tussen de classificaties te maken en door met behulp van data science technieken op basis van bestaande mappings nieuwe mappings te realiseren. Het maken van deze mappings kost veel tijd, en zal opnieuw gemaakt of aangepast moeten worden in het geval van wijzigingen aan een classificatie. Daarom is verkend hoe Natural Language Processing (NLP) technieken dit proces kunnen ondersteunen met automatische of semi-automatische mappings. In dit document zal gerelateerd werk en mogelijke oplossingsrichtingen worden besproken, en een overzicht van uitdagingen waar rekening mee dient gehouden te worden.

Dit rapport is een update van een rapport in 2021, waarbij alleen het werk en inzichten vanuit 2022 zijn toegevoegd. Het werk uit 2021, en dus het gehele rapport van dat jaar is te vinden in de overige secties. Het werk van 2022 is te vinden in sectie 1.3. Dit bevat zowel het werk over het deel van WP1 over hoe de methodologie hoe hybride AI systemen semiautomatisch kunnen leren van bestaande kennisbasen met NLP technieken, alsmede het deel van WP2 over hoe we een kennismodel dynamisch kunnen maken.

1.1 Gerelateerd werk

Neutel & de Boer (2020) hebben een automatische mapping gemaakt tussen de beroepen van ESCO en O*NET. Hierbij is gebruik gemaakt van 5 embedding methodes: fastText labels, descriptions (Bojanowski, 2017), BERT CSL descriptions, BERT mean token descriptions, en SBERT descriptions (Wang & Kuo, 2020). fastText labels zijn word embeddings gemaakt met de fastText library, waarin woord vectors worden gecreëerd zonder context mee te nemen. Bij de descriptions van fastText worden vectors gemaakt van de beschrijvingen van de beroepen. Bij de 3 BERT (Devlin et al., 2018) matching methodes wordt wel context meegenomen op verschillende manieren. Nadat de embeddings zijn gemaakt, wordt elk ESCO beroep vergeleken met een O*NET beroep. De cosine-similarity wordt vervolgens berekend voor elk paar. De performance van deze methoden is niet hoog genoeg voor een direct bruikbare mapping, de mean reciprocal rank van de best presterende methode (SBERT) is 0.503. Deze implementatie kan dus gebruikt worden als inspiratie voor verdere ontwikkeling.

Uit het werk van Neutel & de Boer (2020) blijkt dat de kwaliteit van de embeddings van grote invloed is op de kwaliteit van de uiteindelijke mapping. Bij het mappen van beroepen worden de beschrijvingen ook gebruikt om context mee te nemen. Het mappen van skills (vaardigheden) wordt niet verkend in dit werk, maar is nog niet handmatig uitgevoerd en is daarom de moeite waard om te overwegen. Als we kijken naar een mogelijke mapping tussen de skills van ESCO en CompetentNL zijn er verschillende uitdagingen. Zo zijn de skills van CompetentNL in het Nederlands. ESCO heeft ook een Nederlandse versie, maar dit betekent wel dat Nederlandse embeddings gebruikt moeten worden, die over het algemeen van mindere kwaliteit zijn dan Engelse. Daarnaast hebben skills vaak de vorm van korte phrases met een

noun phrase (NP) gecombineerd met een verb phrase (VP), geen enkel woord maar ook geen zin. Een voorbeeld hiervan is de skill ‘tuinen aanleggen’.

1.2 Oplossingsrichtingen

Mogelijke oplossingen en hun functioneren zullen afhangen van de embeddings gecombineerd met de similarity methode. We kunnen embeddings specifiek trainen op CompetentNL skills en ESCO skills, of een fine-tuned laag toevoegen aan een BERT model. Wat betreft de similariteit berekenen kunnen er ook verschillende methodes worden gebruikt, waarvan de bekendste cosine similarity is (Wang, 2013). Een andere veelbelovende methode is de wordmover’s distance (Kusner et al., 2015). Er kan ook overwogen worden om de skills te filteren op basis van al gematchte beroepen om de taak kleiner te maken. Zo kan er ook gefilterd worden op de soorten skills: er zijn skills, competences, en knowledge. Deze soorten hebben een andere vorm, en door ze te scheiden zullen verkeerde matches tussen die categoriën in ieder geval vervallen. Er zou ook gekeken kunnen worden naar oplossingen die niet naar embeddings kijken, zoals reguliere expressies of synoniemen van wordnet. Doordat de skills deels een NP + VP vorm hebben zou een synoniem van de NP kunnen leiden tot een goede match. Een interessant experiment voor de toekomst is het vergelijken van een niet-statistische methode zoals synoniemen van wordnet of regex met een machine learning aanpak zoals een BERT model.

1.3 Inzichten uit 2022

In 2022 zijn verschillende experimenten gedaan om eerder genoemde oplossingsrichtingen uit te proberen. Daarbij lag de focus op het mappen van skills op elkaar in plaats van beroepen, zoals gebeurde in vorige jaren.

We hebben onderscheid gemaakt in het werk van WP1, dat gaat over een methodologie over het leren van / over verschillende bronnen van informatie, zoals een andere ontologie of een vacature, en het werk van WP2, dat gaat over het dynamisch maken van één kennisbron / ontologie. We hebben daartoe verschillende experimenten gedaan. Deze experimenten worden uitgebreider beschreven in het bijbehorende paper van WP1 & WP2 [de Boer, 2023 – in progress]. Ze omvatten één set experimenten uitgevoerd in het Nederlands – het herkennen van gelijkende skills vanuit verschillende bronnen (WP1) -, en twee sets experimenten in het Engels – het herkennen van gelijkende skills vanuit één bron (WP1/2) en het integreren van nieuwe skills in de ontologie (WP2). Ondanks dat de verschillende WP’s en experimenten verschillende doelen hebben, zijn er gelijkende NLP technieken die ingezet kunnen worden. Deze technieken, zoals de baseline met een simpele Levenshtein distance (LVS), word vectors middels Spacy en transformer modellen BERT(je), JobBERT en XLnet zijn onderling vergeleken met verschillende metrieken, zoals accuracy, top 5 accuracy, Mean Average Precision en Discounted Cumulative Gain.

De resultaten laten zien dat, zoals verwacht, de performance – de uitkomsten van de verschillende metrieken - op het Nederlands een stuk slechter is dan op het Engels. In de analyse blijkt dat dit zowel door de taal komt - de kwaliteit van de verschillende modellen voor de Nederlandse taal vs. de Engelse taal - als dat de

taak van het mappen van verschillende bronnen ten opzichte van binnen dezelfde bron lastiger is. Doordat niet elke ontologie op hetzelfde niveau ageert, is een mapping niet alleen op een volledig synoniem te maken, maar worden ook vagere relaties als ‘narrower’ en ‘broader’ op 1 hoop gegooid. Als we de verschillende NLP technieken vergelijken, blijkt dat afhankelijk van de experimentele setting BERT of Spacy het beste presteert, en dat JobBERT en XLnet niet significant slechter presteren op de verschillende maten. De baseline LVS presteert volgens de verschillende maten in het experiment wel slechter dan de andere methoden, zoals verwacht.

Naast de genoemde resultaten hebben we ook nog de word mover’s distance getest, met een slechtere performance op de verschillende metrieken dan de gebruikte cosine similarity, en in de code een versnelling toegevoegd door middel van het van tevoren uitrekenen van alle embeddings voor de verschillende skills (bij elk van de Transformer en word vector methoden). Waar eerst de code meer dan een dag bezig was om een experiment te runnen duurt het nu nog maar een paar minuten, doordat niet elke keer de embedding opgehaald hoeft te worden uit het model.

In de toekomst willen we 1) uitwerken wanneer een statistische methode nodig is en beter werkt dat de handmatige methode; 2) werken aan een hybride methode die de sterke punten van beide methoden uit 1) combineert; 3) onderscheid maken tussen directe synoniemen en broader en narrower matches; 4) mogelijk kijken naar nieuwe technieken zoals semantische modellen die om kunnen gaan met verschillende typen relaties genoemd in 3)[1] of het maken van een ander type word embeddings[2] ; 5) gebruik maken van de ontologiestructuur (hiërarchie) in plaats van alleen de tekst.

[1] Learning Word Embeddings for Hyponymy with Entailment-Based Distributional Semantics - James Henderson, Oct 2017
[2] Supporting inferences in semantic space: representing words as regions - Katrin Erk, 2009

2 Referenties

- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135-146.
- De Boer, M.H.T., Bakker, R. M. and Burghoorn, M. (2023; in progress). Creating Dynamically Evolving Ontologies: A Use Case from the Labour Market Domain. (to be submitted to Semantics 2023).
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015, June). From word embeddings to document distances. In International conference on machine learning (pp. 957-966). PMLR.
- Neutel, S., & de Boer, M. H. (2021). Towards Automatic Ontology Alignment using BERT. In AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering.
- Wang J. (2013) Pearson Correlation Coefficient. In: Dubitzky W., Wolkenhauer O., Cho KH., Yokota H. (eds) Encyclopedia of Systems Biology. Springer, New York, NY. https://doi.org/10.1007/978-1-4419-9863-7_372
- Wang, B., & Kuo, C. C. J. (2020). Sbert-wk: A sentence embedding method by dissecting bert-based word models. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, 2146-2