Linköpings universitet | Institutionen för Datavetenskap
Kandidatuppsats, 18 hp | Kognitionsvetenskap
Vårtermin 2019 | LIU-IDA/KOGVET-G--19/001--SE

Word2vec2syn

Synonymidentifiering med Word2vec

Tove Pettersson

Handledare: Robert Eklund

Examinator: Arne Jönsson

Uppdragsgivare: Fodina Language Technology AB

Sammanfattning

Inom NLP (eng. natural language processing) är synonymidentifiering en av de språkvetenskapliga utmaningarna som många antar. Fodina Language Technology AB är ett företag som skapat ett verktyg, Termograph, ämnad att samla termer inom företag och hålla den interna språkanvändningen konsekvent. En metodkombination bestående av språkteknologiska strategier utgör synonymidentifieringen och Fodina önskar ett större täckningsområde samt mer dynamik i framtagningsprocessen. Därav syftade detta arbete till att ta fram en ny metod, utöver metodkombinationen, för just synonymidentifiering. En färdigtränad Word2vec-modell användes och den inbyggda funktionen för cosinuslikheten användes för att få fram synonymer och skapa kluster. Modellen validerades, testades och utvärderades i förhållande till metodkombinationen. Valideringen visade att modellen skattade inom ett rimligt mänskligt spann i genomsnitt 60,30 % av gångerna och Spearmans korrelation visade på en signifikant stark korrelation. Testningen visade att 32 % av de bearbetade klustren innehöll matchande synonymförslag. Utvärderingen visade att i de fall som förslagen inte matchade så var modellens synonymförslag korrekta i 5,73 % av fallen jämfört med 3,07 % för metodkombinationen. Den interna reliabiliteten för utvärderarna visade på en befintlig men svag enighet, Fleiss Kappa = 0,19, CI(0,06, 0,33). Trots viss osäkerhet i resultaten påvisas ändå möjligheter för vidare användning av word2vec-modeller inom Fodinas synonymidentifiering.

Nyckelord: Word2vec, synonymidentifiering, vektorrymdsmodell, ordvektorer, cosinuslikhet

Abstract

One of the main challenges in the field of natural language processing (NLP) is synonym identification. Fodina Language Technology AB is the company behind the tool, Termograph, that aims to collect terms and provide a consistent language within companies. A combination of multiple methods from the field of language technology constitutes the synonym identification and Fodina would like to improve the area of coverage and increase the dynamics of the working process. The focus of this thesis was therefore to evaluate a new method for synonym identification beyond the already used combination. Initially a trained Word2vec model was used and for the synonym identification the built-in-function for cosine similarity was applied in order to create clusters. The model was validated, tested and evaluated relative to the combination. The validation implicated that the model made estimations within a fair human-based range in an average of 60.30% and Spearmans correlation indicated a strong significant correlation. The testing showed that 32% of the processed synonym clusters contained matching synonym suggestions. The evaluation showed that the synonym suggestions from the model was correct in 5.73% of all cases compared to 3.07% for the combination in the cases where the clusters did not match. The interrater reliability indicated a slight agreement, Fleiss' Kappa = 0.19, CI(0.06, 0.33). Despite uncertainty in the results, opportunities for further use of Word2vec-models within Fodina's synonym identification are nevertheless demonstrated.

Key words: Word2vec, synonym identification, vector space model, word vectors, cosine similarity

Förord

För det slutgiltiga resultatet och uppkomsten av detta kandidatarbete har jag flertal involverade vars bidrag varit av stor betydelse. Först och främst vill jag därför rikta ett stort tack till Fodina som initierat projektet och därmed gjort detta kandidatarbetet möjligt. Jag vill även tacka för all hjälp och stöd som ni givit mig på vägen. Även seminariegruppen, inkluderat handledare och examinator, har under arbetets gång kommit med givande förslag som berikat detta arbete. Slutligen vill jag även tacka min pappa som under hela arbetets gång agerat bollplank och visat en enorm hjälpsamhet och ett stort engagemang. Stort Tack!

Linköping, 7 juni 2019
Tove Pettersson

Innehållsförteckning

1. Introduktion

Att analysera och bearbeta språk är mycket resurskrävande både vad gäller tid men också kognitivt i form av mentala resurser. I en tid av digitalisering är det viktigt att utnyttja den teknik och de resurser som finns tillgängliga. Detta tekniska paradigm tillsammans med människors ständiga strävan efter att underlätta och förenkla den egna belastningen samt den ökande mängden tillgänglig text på internet kan tillsammans utgöra bidragande faktorer till det växande intresset inom språkteknologi (eng. Language Technology). Språkteknologi är ett tvärvetenskapligt område som har sin grund inom lingvistik och datalogi och syftar främst till att analysera eller på annat sätt bearbeta det naturliga språket med hjälp av datavetenskapliga modeller eller program. Processen kallas NLP som är en akronym från engelskans *Natural Language Processing*.

Språk är komplext för människan utifrån flera aspekter, dels består dess uppbyggnad av flera olika dimensioner av relationer såsom syntaktiska, semantiska och morfologiska som erbjuder olika sätt att se på ett språk och samhörigheter inom det. Dessutom är det som sägs ofta starkt knutet till den kontext det sägs i vilket gör att dess tolkning kan bli helt olika om ett ord tas ifrån sin ursprungliga kontext.

Dessa starka band mellan ord och kontext är något som med fördel kan utnyttjas inom olika former av NLP. En typ av NLP där detta kan göras är inom synonymidentifiering. En dator kan inte tolka ett ords semantiska mening så som människor kan vilket gör detta till en problematik. Det har dock visat sig finnas ett flertal metoder som med hänsyn till just ett ords kontext kan bevara en del semantiska relationer som finns i meningar.

Genom att utforma matriser för ord och dess förekomster utifrån en viss kontext så kan varje ord ersättas med data som är mer hanterbar för datorer, det vill säga siffror. På så vis skapas ordvektorer inom så kallade vektorrymdsmodeller som matematiska beräkningar kan appliceras på för att återkalla de semantiska relationerna som påträffats. En välutformad modell kan i sin tur användas som ett kompletterande verktyg för att reducera brister i människors språkliga färdigheter eller för effektivisering.

1.1 Syfte

Fodina Language Technology AB är ett språkteknologiskt företag som skapat ett verktyg för att samla termer inom företag som de kallar för Termograph. Termograph är ett interaktivt verktyg som används för att granska, modifiera och publicera terminologikoncept. Genom att bland annat ta fram synonymer kan ersättande och föredragna ord som ett specifikt företag ämnar använda föreslås. Synonymidentifieringen utgör det första steget och behandlas sedan ytterligare i Termograph. Synonymidentifieringen är idag baserad på en kombination av flera metoder inom språkteknologi och kommer därav vidare benämnas som metodkombinationen. Som komplement till metodkombinationen önskar Fodina en ytterligare metod.

Den efterfrågade metoden bör vara mer dynamisk att arbeta med och ha större räckvidd, alltså få fram andra synonymer än den nuvarande metodkombinationen. Därmed är syftet med detta kandidatarbete att implementera en metod för synonymidentifiering som inte redan utgörs av metodkombinationen. Detta arbete anses vara ett första steg att testa och utvärdera möjligheterna som ordvektorer i en vektorrymdsmodell, mer specifikt en Word2vec-modell kan erbjuda för Termograph.

Anledningen till valet av ordvektorer baseras på det breda användningsområde som utgörs av dessa (Nurifan, Sarno och Wahyuni, 2018; Nguyen, Schulte im Walde och Vu, 2017). Utöver synonymidentifiering kan vektorrymdsmodeller även tränas för att identifiera exempelvis homonymer och antonymer vilket kan komma till användning för olika språkliga specificeringar vid ett senare skede men också för att öka chanserna att uppnå det efterfrågade ökade täckningsområdet.

1.1.1 Problematik.

Generellt sett för många metoder inom synonymidentifiering och så även för metodkombinationen så finns det framförallt tre olika framstående utmaningar. Detta rör sig om problematiken med homonymer, hyperonymer och hyponymer.

Homonymer är ord som ser likadana ut men som har olika betydelser, dessa ord är alltså endast tolkningsbara utifrån den kontext de används i (Saeed, 2015). Detta gör att många metoder inte gör skillnad på homonyma ord och alltså placerar dessa i ett och samma synonymkluster. Vilket gör att exempelvis ordet ”plan” skulle ge ett synonymkluster som både innehåller synonymen ”platt” men också ”flyg” som i sin tur är två helt olika ord och därav bör tillhöra två separata synonymkluster. Även ord inom samma ordklass kan vara homonymer vilket ökar svårigheten att skilja dem åt.

Problemet med hyperonymer och hyponymer, det vill säga, över- och underordnade begrepp handlar om ord såsom ”frukt” och ”äpple” där ordet äpple visserligen är en typ av frukt men inte synonymt med det (Saeed, 2015). Frukt är alltså en hyperonym till ordet äpple som i sin tur är en hyponym till ordet frukt.

1.1.2 Problemformulering.

Studien innebär att identifiera synonymer ur en korpus med hjälp av en tränad word2vec-modell i syfte att komplettera metodkombinationen och slutligen Termograph. Av denna anledning kommer modellen att fristående valideras, därefter testas i syfte att jämföras med metodkombinationens befintliga resultat och slutligen utvärderas manuellt för att närmare utreda hur synonymförslagen skiljer sig åt, detta då ingen guldstandard, det vill säga facit, finns tillgänglig.

1.2 Avgränsning

En avgränsning kommer att göras dels för att säkerställa att arbetet fullföljs inom de angivna tidsramarna samt för att fokus inom detta arbete är att ta fram en ny metod utöver metodkombinationen och inte att hantera befintliga fel. Den enda anpassning av metoden i detta syfte som kommer att göras är att helt utesluta WordNet som varit en stor källa för metodkombinationen. Detta för att undgå att den vanliga problematiken ska överföras även till modellen. Detta gäller även om det inte är klarlagt att felen uppstått på grund av WordNet. Ingen vidare anpassning eller justering av metoden i syfte att undgå de mest frekvent förekommande felen hos metodkombinationen kommer att göras. Homonymer, hyperonymer och hyponymer kommer alltså inte att hanteras då detta skulle ge upphov till ytterligare problematiker och därmed bör dessa istället hanteras separat.

1.3 Frågeställning

För att uppfylla syftet med studien så kommer följande frågeställningar att besvaras:

- Kan en Word2vec-modell användas för att ta fram rimliga synonymförslag?
- Kan en Word2vec-modell erbjuda ett större täckningsområde på så vis att modellen får fram andra synonymer än metodkombinationen?
- Kan en Word2vec-modell användas som en metodik för metodkombinationens synonymidentifiering?

2. Bakgrund

I detta kapitel presenteras bakgrundsinformation i form av teori och tidigare studier som anses relevant för vidare läsning och förståelse av denna rapport.

2.1 Semantiska relationer

Semantik handlar om meningen eller tolkningen av ord. Språk består av en mängd semantiska relationer som definierar hur ord hör ihop utifrån dess betydelse. För förståelse och korrekt användning av språk utgör semantiska relationer en grundläggande byggsten (Saeed, 2015). Det har visat sig att ord som är nära relaterade med varandra semantiskt tolkas på liknande sätt i hjärnan vilket bland annat framkommit genom studier och observationer av människor som lider av afasi. Afasi är en sjukdom som medför svårigheter att bilda korrekta meningar då människor som lider av detta tenderar att göra fel ordval. Detta kan medföra att meningarna inte blir semantiskt tolkningsbara. Det har dock visat sig att dessa felaktiga ordval ofta har en nära semantisk relation till det korrekta ordet.

En typ av semantisk relation är synonymer vars uppkomst föreslås härstamma från olika språk, dialekter eller grad av formalitet i språket (Saeed, 2015). Definitionen av synonymer är ord som har liknande betydelse som varandra och inte nödvändigtvis exakt samma betydelse. Detta innebär vidare att orden inte behöver vara utbytbara i alla kontexter, så kallade kompletta synonymer (eng. complete synonymy). Det hävdas att kompletta synonymer är mycket sällsynta om de överhuvudtaget existerar i det naturliga språket och därmed räcker det med att två ord refererar till samma objekt för att anses vara synonymer, så kallade delvisa synonymer (eng. partially synonymy). Det som hädanefter kommer att betecknas synonymer inkluderar därmed alla typer av dessa, det vill säga, ord som refererar till samma objekt oavsett om de alltid är utbytbara i varandras kontexter eller inte.

2.2 Neurala nätverk

Maskininlärning är ett brett begrepp för flera inlärningsmetoder som har gemensamt att de drivs och baseras på återkoppling (eng. feedback) (Russel & Norvig, 2016). En modell tränas upp och resultatet justeras sedan utifrån den återkoppling som modellen får. Det finns framförallt två olika typer av maskininlärning som grundar sig i om det finns en tillhörande guldstandard för den specifika inlärningsuppgiften eller inte. När det finns en guldstandard kallas det för övervakad inlärning (eng. supervised learning). Denna typ av inlärning drivs genom att para ihop input med ett korrekt output som finns givet i guldstandarden. Vid oövervakad inlärning

(eng. unsupervised learning) finns ingen direkt guldstandard så modellen lär sig genom att hitta mönster i den input som finns. En vanlig uppgift för oövervakad inlärning handlar om klustring, där input grupperas utifrån tillhörigheter. Denna klustring kan exempelvis utgöras av ord som delar någon form av likhet såsom synonymer och dess semantiska relationer.

Maskininlärning är ett brett område som innefattar flera olika metoder. För detta arbete är det dock framförallt en grundläggande förståelse för neurala nätverk, en typ av maskininlärning, som är av stor vikt. Övriga delar av maskininlärning ligger utanför ramen av detta arbete.

Neurala nätverk beskriver den matematiska tolkning som gjorts av hjärnan och dess funktionalitet (Russell & Norvig, 2016). Den teori som ligger till grund för neurala nätverk är att all aktivitet i hjärnan bedrivs av cellerna inuti den. Dessa hjärnceller kallas neuroner och den simpla matematiska tolkningen av dessa är att de avfyrar vid detektion av specifika mönster vilket har inspirerat till utvecklingen av de neurala nätverken. Ett neuralt nätverk består av ett flertal sammankopplade enheter som tillsammans utgör inputlager, dolt lager samt outputlager. Det neurala nätverkets funktionalitet beror på neuronernas detekterande egenskaper. Mellan enheterna propageras aktiviteten via direkta länkar vars styrka och riktning indikeras av en så kallad vikt. All indata till varje enhet summeras med hjälp av en aktiveringsfunktion som genererar utdata. Beroende på hur stark den totala aktiveringen i enheten är i förhållande till tröskelvärdet så justeras utdata där 1 brukar indikera en avfyrning för neuronen medan 0 indikerar att aktiveringen inte överskred tröskelvärdet. Det främsta användningsområdet för neurala nätverk är inom inlärning.

2.3 Vektorer

En vektor är en matematisk enhet som används för illustrationer och beräkningar av kvantiteter som har både mätetal och riktning, exempelvis en hastighet (Andersson, Grennberg, Hedberg, Näslund, Persson & Sydow, 1999). Längden på vektorn bestäms utifrån värdet på mätetalet och vektorn ska peka åt riktningens håll. Den riktade sträckan är definierad inom sträckans fotpunkt till dess spets. När två vektorer har samma riktning och samma längd indikerar detta att den riktade sträckan som beskrivs av vektorerna är lika. Inom neurala nätverk kan vektorer användas för att beskriva vikterna i nätverket.

2.3.1 Cosinuslikhet.

Cosinuslikhet (eng. cosine similarity) kan användas i syfte att beräkna likheter mellan ordvektorer (Lei, Si, Wen & Shen, 2017). Cosinuslikheten beräknar riktningen på en vektor genom att använda vinkelavståndet mellan två ordvektorer. Denna riktning avslöjar närheten i vektorrymden mellan de specifika ordvektorerna och brukar användas för att avgöra hur lika orden är där lika riktning även indikerar lika ord. Cosinuslikheten tar inte hänsyn till magnituden vilket ofta anses vara fördelaktigt vid just ordvektorer då magnituden inte är av betydande roll. Värdet som tas fram för cosinuslikheten är mellan -1 och 1 där -1 ämnar indikera antonymer, det vill säga motsatser och 1 ämnar indikera synonymer. Vidare innebär detta att ju mindre vinkel mellan ordvektorerna desto närmare 1 blir cosinuslikheten. Eftersom syftet med cosinuslikheten ofta är att jämföra likheter så används värdet främst i positiva vektorrymden och därav genereras ett värde mellan 0 och 1. Där 0 indikerar att orden är orelaterade till varandra.

2.3.2 Euklidiskt avstånd.

Det Euklidiska avståndet är det så kallade normalavståndet mellan två ordvektorer och desto mindre värde ju mer lika är orden (Lei, Si, Wen & Shen, 2017). Värdet beräknas genom Pythagoras sats där den räta vinkeln placeras mellan de två ordvektorerna och därmed placerar hypotenusan som en linje från vektor 1 till vektor 2. Eftersom det euklidiska avståndet mäter sträckan mellan två ordvektorer så har magnituden i vektorrymden en betydelse till skillnad från när cosinuslikheten används. Det euklidiska avståndet kan dock normaliseras för att eliminera magnitudens betydelse och därav även eliminera betydelsen av frekvensen av ord. Det normaliserade euklidiska avståndet blir därmed mycket likt det värde som framtas genom cosinuslikheten.

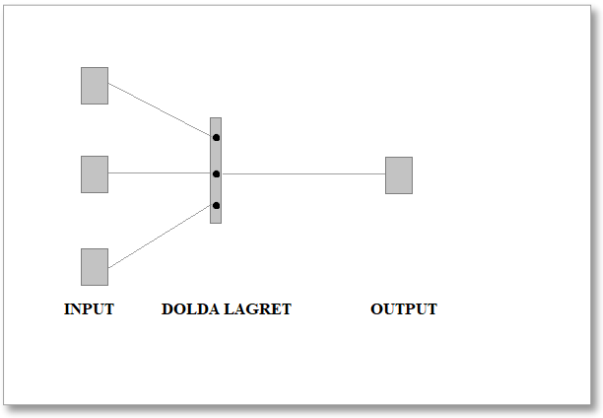
2.4 Word2vec

Word2vec är en vektorrymdsmodell som grundat sin arkitektur utifrån neurala nätverk som fungerar genom att hela tiden förmedla fel bakåt i modellen genom bakåtpropagering (eng. backpropagation). Detta ger viktjusteringar i det dolda lagret som i sin tur representeras av en ordvektor i vektorrymden (Rong, 2014). Den grundläggande teorin bakom Word2vec är att ett ords betydelse ska kunna tolkas utifrån den kontext, det vill säga mening, som ordet befinner sig i. Detta eftersom liknande ord ofta förekommer i liknande kontext (Mikolov, Chen, Corrado, & Dean, 2013). Modellen skapar en matris till varje ord som baseras på de kontextuella egenskaperna, det vill säga de ord som vanligen förekommer tillsammans i en specifik mening.

Genom att tilldela ord en vektorrepresentation genom denna matris så kan semantiska relationer mellan ord bibehållas. För att ta fram vektorrepresentationerna används en av följande grundläggande metoder, CBOW (eng. continuous bag of words) eller SG (eng. skip-gram).

2.4.1 CBOW.

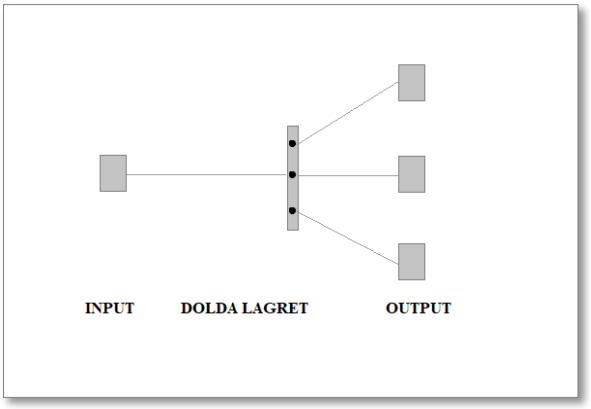
CBOW är en förhållandevis simpel metod som därav innebär att den även har en kortare bearbetningstid än andra metoder. Metoden är mest lämplig att använda vid större datamängder då den vid för små datamängder visats ha stora brister (Lai, Liu, He & Zhao, 2016). CBOW förutser ett ord (eng. target word) givet den tillhörande kontext för det specifika ordet (Rong, 2014). Modellen beräknar vikterna för varje ord utifrån ett medelvärde av alla ordvektorer i meningen. I Figur 1 nedan illustreras en enkel modell av metodens uppbyggnad.



Figur 1. CBOW-modell där flera input endast genererar ett output via det dolda lagret i modellen. Input kan vara en mening (flera ord) och output ett ord.

2.4.2 SG.

SG gör det motsatta mot CBOW och förutser istället en kontext utifrån ett specifikt ord (Rong, 2014). Figur 2 illustrerar metodens uppbyggnad. Metoden är baserad på n-gram såsom exempelvis ordpar av två (bigram) eller tre (trigram) och så vidare. SG är en förhållandevis komplex metod som därav innebär att den även har en längre bearbetningstid än mer simpla metoder. Metoden är användbar både vid stora och små datamängder (Lai, Liu, He & Zhao, 2016). SG har även visat sig vara framgångsrik för att identifiera semantiska relationer (Mikolov, Chen, Corrado & Dean, 2013).



Figur 2. SG-modell där varje enskilt input genererar flera output via det dolda lagret i modellen. Input kan vara ett ord och output dess mening (flera ord).

2.4.3 Vanliga problem.

Ett vanligt förekommande problem som uppstår i samband med ordvektorer som metod är att modellen kan komma att missta antonymer, såsom ”varmt/kallt” eller ”glad/ledsen”, för synonymer. Detta eftersom metoden är kontextbaserad vilket leder till att ord som ofta förekommer i liknande kontext därmed tolkas som synonymer (Scheible, Schulte im Walde & Springorum, 2013). Att byta ut ett ord i en mening mot sin antonym ändrar meningens semantiska betydelse men är syntaktiskt korrekt, därav denna problematik. Även homonymer, hyponymer och hyperonymer kan utgöra ett problem. Dock kan samtliga av dessa problem delvis elimineras med en tillräckligt stor datamängd eftersom prestandan hos Word2vec-modeller överlag förbättras desto större datamängd modellen tränats på (Sahlgren & Lenci, 2016).

2.4.4 Parametrar.

Det finns många olika parametrar inbyggda i word2vec som är av betydande roll både för träning men därmed också för modellens totala prestanda (Dridi, Gaber, Azad & Bhogal, 2018; Tezgider, Yıldız & Aydın, 2018; Ganesan, 2018). Några av dessa är *size*, *window* och *min_count*. *Size* är alltså storleken på vektorn vilket innebär att vid små datamängder bör denna inte vara för stor (Ganesan, 2018). Rekommenderat är att hålla denna mellan ett värde på 100-150 beroende på hur stor datamängd som används. Det finns även resultat av parameteroptimeringar som visar på att inställningar av *size* inom spannet av 150 – 300 är att rekommendera (Tezgider, Yıldız, & Aydın, 2018).

Parametern *window* definierar hur långt avståndet mellan det specifika ordet till sitt kontextord får vara (Ganesan, 2018). Om denna sätts till exempelvis 10 så innebär det att det får vara som mest 10 ord mellan för att räknas som kontext till det specifika ordet. Teoretiskt sett är det fördelaktigt att sätta detta värde till så lågt som möjligt men justeringarna bör utföras med hänsyn till storleken på datamängden.

Parametern *min_count* innebär hur många gånger ett ord behöver förekomma i en korpus för att få vara delaktig i bearbetningen (Ganesan, 2018). Denna bör sättas till minst 2 för att få bort sällsynta ord som endast förekommer en gång men samtidigt säkerställa en så stor vokabulär som möjligt.

Utöver parametrarna har det även visat sig vara av stor betydelse att den data som används är domänspecifik och att detta i vissa fall till och med kan vara av större betydelse än att ha en riktigt stor datamängd (Lai, Liu, He & Zhao, 2016). Dock har det även visat sig att ordvektorer kräver en stor datamängd för att uppnå en god övergripande prestanda och för att uppnå sin fulla kapacitet (Sahlgren & Lenci, 2016).

2.5 Tidigare studier

Många tidigare studier har använt sig av modeller såsom Word2vec för att identifiera olika typer av semantiska relationer mellan ord (Zhou, Fu, Qiu, Zhang & Liu, 2017; Lee & Lee, 2018). Det är även vanligt att använda just cosinuslikheten för att ta fram dessa likheter, framförallt för identifiering av synonymer (Lei, Si, Wen & Shen, 2017). Vektorrymdsmodeller har även ett bredare användningsområde som sträcker sig utöver synonymidentifiering vilket kan vara fördelaktigt för arbete med språk där en problematik också ofta medför flera andra problem.

2.5.1 Jämförelse av vektorrymdsmodeller.

Det finns olika modeller och olika varianter av dessa i sin tur vilket gör att varje modell är optimerad för sina specifika områden. De överlägset mest använda modellerna överlag är Word2vec och GloVe (eng. Global vectors for word representation). Berardi, Esuli och Marcheggiani (2015) utförde en studie där de undersökte skillnaderna mellan dessa modeller. En Word2vec-modell och en GloVe-modell tränades på två datamängder vardera i syfte att utöver skillnader mellan de olika modellerna även urskilja skillnader genererade av träningsdata inom modellerna. En engelsk datamängd användes samt en italiensk i syfte att framhäva de språkliga skillnaderna morfologiskt mellan engelska och italienska. Word2vec-modellen presterade överlägset bäst både på italienska men framförallt på engelska. Noggrannheten (eng.

accuracy) mättes till 47 % för italienska respektive 60 % för engelska. För specifikt semantiska uppgifter presterade word2vec-modellen bättre än GloVe-modellen med en noggrannhet på 48,81 % av Word2vec-modellen respektive 21,33 % av GloVe-modellen då båda var tränade på data från Wikipedia. Resultatet av denna studie indikerar därmed att en Word2vec-modell kan vara fördelaktig i just semantiska uppgifter, såsom synonymidentifiering.

2.5.2 Synonymidentifiering med vektorrymdsmodeller.

Kompletta synonymer är utbytbara med varandra i dess kontext vilket inte är fallet för alla synonymer (Saeed, 2015). Dock indikerar detta att ju mer synonyma ord är med varandra desto oftare bör fallet vara att orden även är utbytbara med varandra vilket vidare innebär att metoder som utgår från just kontexten är en mycket vanlig och framgångsrik metod vid identifiering och extrahering av synonymer, däribland vektorrymdsmodeller (Lee & Lee, 2018; Zhou, Fu, Qiu, Zhang & Liu, 2017). Lee och Lee (2018) utförde en studie i syfte att designa ett diagnostiseringssystem där systemet skulle kunna diagnostisera patienter med hjälp av en symptombeskrivning, en beskrivning i naturligt språk. För att kunna göra detta på ett framgångsrikt sätt så krävdes att systemet kunde identifiera synonymer för att därmed förstå beskrivningen den tillhandahölls med. Detta gjordes med hjälp av en Word2vec-modell som ensamt användes för att lösa problem med synonymer som symptombeskrivningarna eventuellt innehöll. Diagnostiseringssystemets totala prestanda förbättrades då Word2vec användes som synonymhanterare. Detta resultat indikerar att en Word2vec-modell kan utgöra en bra metod för just synonymidentifiering.

Även Zhou, Fu, Qiu, Zhang och Liu (2017) utförde en liknande studie i syfte att ta fram de semantiska relationerna mellan kinesiska medicinska termer. En SG-modell jämfördes med en CBOW-modell för att undersöka vilken av metoderna som verkade mest pålitlig för det specifika ändamålet. Resultatet visade att genom att använda sig av en SG-modell och med hjälp av denna skapa vektorer av de medicinska termerna så uppnåddes en ökning av dels den totala prestandan samt en ökning på 15 % i noggrannhet.

2.5.3 Synonymidentifiering med cosinuslikhet.

Ett flertal studier har utförts i syfte att identifiera synonymer där bland annat cosinuslikhet använts för att ta fram dessa synonymer. Zhang, Li och Wang (2017) utförde en studie i syfte att extrahera synonymer ur en korpus. En Word2vec-modell med SG-metoden användes för att skapa vektorer och synonymerna extraherades sedan med hjälp av cosinuslikheten. Synonymkluster skapades baserat på det beräknade värdet av cosinuslikheten

som användes för att skapa ett spektrum i vektorrymden, så kallad spektrumbaserad klustring (eng. spectral clustering). En precision (eng. precision) på 80,8 %, återkallning (eng. recall) på 74,4 % samt ett F1-värde på 0,775 uppnåddes då denna spektrumbaserade metod användes, vilket var betydligt bättre än studiens baslinje (eng. baseline), en metod kallad K-means som uppnådde en precision på 27,9 %, återkallning på 47,3 % samt ett F1-värde på 0,351.

Lei, Si, Wen och Shen (2017) utförde en studie för att klassificera om två kinesiska medicinska termer var synonymer eller inte till detta användes en huvudmetod baserad på stödvektormaskiner (eng. Support Vector Machine).

En stödvektormaskin är en modell anpassad för övervakad inlärning och baseras på olika inlärningsalgoritmer främst avsedda för just klassificering men kan också användas vid regressionsanalys.

I studien användes både kinesiska men också engelska termer och de utgick från 13 olika typer av algoritmer för att avgöra den slutgiltiga klassificeringen. Av dessa 13 funktioner hade tre anknytning till ordvektorer, sex stycken hade anknytning till ordnivåer, två stycken med semantisk anknytning och även två stycken specifika för det kinesiska språket.

För att ta fram synonymer användes utöver cosinuslikheten, som ytterligare säkerhet, även det euklidiska avståndet som en kompletterande beräkning. Vidare användes utöver cosinuslikheten och det euklidiska avståndet även andra metoder som alla värderades både individuellt samt i kombination med varandra och det visade sig att den bästa kombinationen innefattades av bland annat cosinuslikheten då en precision på 97,37 %, återkallning på 96,00 % samt ett F1-värde på 0,97 uppnåddes.

2.5.4 Homonymidentifiering med cosinuslikhet.

Homonymer är en klassisk utmaning inom urskiljning av ords betydelser (eng. word sence disambiguation [WSD]) och även inom detta område kan cosinuslikheten appliceras. Nurifan, Sarno och Wahyuni (2018) använde sig av ordvektorer i en komplex studie för just WSD med syfte att urskilja homonympar vilket utfördes med hjälp av en word2vec-modell samt genom framtagning av cosinuslikheten. Två olika betydelser av ett homonympar användes som input och därmed genererades två olika korpusar, en för vardera betydelse. För att hantera problemet med ovanliga ord, det vill säga ord som inte existerat i träningsmängden och som därmed är osedda för modellen så användes Lesk-algoritmer och Wu Palmers likheter (eng. Wu Palmer similarity). Denna hantering ökade även måttet på noggrannhet som resulterade i 85,51 % som visade sig vara en ökning med 8,02 % då skillnader mellan homonympar togs hänsyn till.

2.5.5 Mänsklig synonymidentifiering.

Synonymidentifiering är en komplex uppgift och även människor har ibland svårigheter att enas om likheter mellan ord. *The WordSimilarity-353 Test Collection* är en datamängd bestående av en samling engelska termer med tillhörande mänsklig skattningssiffra av ordparens likheter (Finkelstein, Gabrilovich, Matias, Rivlin, Solan, Wolfman & Ruppin, 2001). Datamängden är framtagna i syfte att redogöra för semantiska likheter mellan ord, enligt en mänsklig uppfattning. Samlingen av termer (testsamling1) presenterades parvis för samtliga testdeltagare (N =13) som ombads att uppskatta ordens likheter genom att ange en siffra mellan 0 till 10, där 0 indikerade att orden var långt ifrån relaterade och 10 indikerade en nära relation. Samma procedur utfördes för testsamling2 och de andra deltagarna (N = 16). Studiens resultat från de båda testssamlingarna finns bevarade i datamängden med tillhörande mänsklig uppskattning av hur lika termerna anses vara. WordSimilarity-353 Test Collection har använts som validering och test vid ett flertal studier (Milne & Witten, 2008; Strube & Ponzetto, 2006). Däribland för att undgå överanpassning (eng. overfitting) vid träning av modeller men också för att evaluera mått för semantiska relationer. Vid det sistnämna har Pearsons korrelation beräknats för att jämföra de semantiska relationernas likheter.

Vidare för att avgöra hur väl människors uppfattning av synonymer eller andra kategoriseringar inom språk faktiskt representerar verkligheten kan utvärderarens interna reliabilitet (eng. interrater reliability) beräknas (McHugh, 2012). Detta värde presenteras av ett så kallat Kappa-värde och är vanligt just inom språkteknologi eftersom språk kan vara mycket tvetydigt. Denna tvetydighet orsakar en stor osäkerhet vilket medför svårigheter att erhålla en pålitlig guldstandard. Syftet med att beräkna Kappa-värdet är att stärka trovärdigheten av testresultaten eftersom mänskliga utvärderingar oftast involverar viss mängd osäkerhet eller slump.

3. Metod

För en relevant inblick av metoden behandlas i detta kapitel tre olika deluppgifter som utgjordes av validering, testning och utvärdering med tillhörande data och beskrivning av utförandet.

3.1 Data

Tre olika datamängder var relevanta för genomförandet. Dessa bestod av träningsdata, valideringsdata samt testdata.

3.1.1 Träningsdata.

Den data som modellen tränats på är den kompletta texten från Engelska Wikipedias databasdump från februari 2017 (Nordic Language Processing Laboratory [NLPL], u.å; Wikimedia Foundations, 2017). Detta utgjorde en korpusstorlek på ca 2,3 miljarder tecken och en vokabulär innehållande 296 630 unika lemman. Anledningen till att en modell tränad på Wikipedia-data användes var främst för att erhålla en stor träningsdatamängd men också för att undvika problem med ord som inte finns i modellens vokabulär, så kallade OOV-ord (eng. Out of vocabulary). Eventuella flerordstermer, det vill säga termer bestående av flera ord, som ingick i träningsdata hanterades inte av modellen vilket vidare innebar att inga sådana termer ingick i modellens vokabulär.

3.1.2 Valideringsdata.

Den datamängd som användes i syfte att validera modellen, så kallad valideringsdata, var *The WordSimilarity-353 Test Collection* (Finkelstein, Gabrilovich, Matias, Rivlin, Solan, Wolfman & Ruppin, 2001). En engelsk datamängd som bestod av termer i par med tillhörande mänskliga uppskattningar av hur lika dessa termer anses vara. Datamängden bestod av totalt två testsamlingar och uppskattningarna presenterades dels med ett medelvärde och dels med alla enskilda uppskattningar från samtliga deltagare (N = 13 respektive N = 16). Den totala datamängden bestod av totalt 353 ord varav testsamling1 innehöll 153 ordpar och testsamling2 innehöll 200 ordpar. Nedan presenteras strukturen för varje rad i testsamlingarna.

ord 1,ord 2, uppskattningar: (medelvärde),1,2,3,4,5,6,7,8,9,10,..

3.1.3 Testdata.

Testdata har framtagits genom att först låta metodkombinationen skapa synonymkluster från en engelsk korpus. Den korpus som bearbetades var extraherad från Wikipedia och baserad på tre domänspecifika teman; *Aviation, Telecommunication* och *Trucks*. Dessa användes som

huvudlänkar hos Wikipedia varav samtliga texter extraherades inklusive alla länkade sidor från huvudsidorna. Detta gav en total korpusstorlek bestående av 4 738 092 ord och 25 178 679 tecken. De synonymkluster som skapades av metodkombinationen utifrån korpusen bestod av 59 497 ord och utgjorde alltså den fulla mängden testdata innan bearbetning. Klusterstorleken var av stor variation och löpte mellan 2 till 218 ord per kluster (M = 3,99, SD = 8,11) vilket dels berodde på att datamängdens främsta syfte var att ha en hög återkallning för att inte missa synonymförslag. Vidare bestod synonymklustren av flera sammankopplade synonympar som gemensamt framtagits av metodkombinationen vilket även kan ha utgjort en faktor till de stora klusterstorlekarna.

För att testdata skulle gå att applicera på själva testningen så utfördes en bearbetning som innebar omvandling av versaler till gemener och utsortering av dubbletter av ord. Denna bearbetning preciseras ytterligare under avsnitt 3.2 Utförande. Efter bearbetningen varierade klusterstorleken mellan 1 och 189 ord per kluster (M =3,43, SD = 6,78).

3.2 Utförande

Utförandet bestod av fyra olika moment vars resultat tillsammans användes för att besvara samtliga frågeställningarna. Dessa moment var träning, validering, testning och manuell utvärdering.

3.2.1 Träning.

En förtränad modell användes (Fares, Kutuzov, Oepen, Velldal, 2017). Modellen hade tränats på hela mängden beskriven i 3.1.1 träningsdata. Träningsdata var först bearbetad genom NER-taggning (eng. named entity recognition tagging) med hjälp av Stanford Core NLP v. 3.6.0, borttagning av stoppord med hjälp av NLTK samt lemmatisering. Modellen tränades med metoden SG och parametern *window* hade inställningen 5 och *size* var lika med 300. Vid träningstillfället genomfördes totalt 5 iterationer. Övriga parametrar behöll sitt förinställda värde vilket bland annat innebar att parametern *min_count* var inställd på 5 under träningen.

3.2.2 Validering.

Valideringen utfördes i syfte att besvara frågeställningen: *kan en Word2vec-modell användas för att ta fram rimliga synonymförslag?* Valideringen utfördes med Python 3.7 genom att jämföra modellens framtagna cosinuslikhet med valideringsdata från *The WordSimilarity-353 Test Collection*. Ordparen från valideringsdatamängden skickades in i modellen och

cosinuslikheten mellan orden beräknades. Värdet på cosinuslikheten normaliserades genom att multipliceras med 10 för att anpassas till den skala som de mänskliga skattningarna gjorts på, det vill säga 0 - 10. Det normaliserade värdet för cosinuslikheten (*norm_cos_sim*) jämfördes sedan med de mänskliga skattningarna i valideringsdatamängden. Om *norm_cos_sim* låg inom spannet av det lägst skattade värdet och det högst skattade värdet så innebar detta att modellens skattning hamnade inom spannet av vad en människa skulle kunnat skatta, vilket därmed bör anses vara rimligt. Därmed ansågs modellens framtagna värde som rimligt och också godkänt. Annars ansågs det vara felaktigt.

I tillägg till denna validering utfördes även en korrelation mellan de mänskliga skattningarna och cosinuslikheten framtaget av modellen. Eftersom den data som användes var på ordinalskalenivå så användes Spearmans korrelationskoefficient som applicerades direkt på datamängden utan att först normalisera.

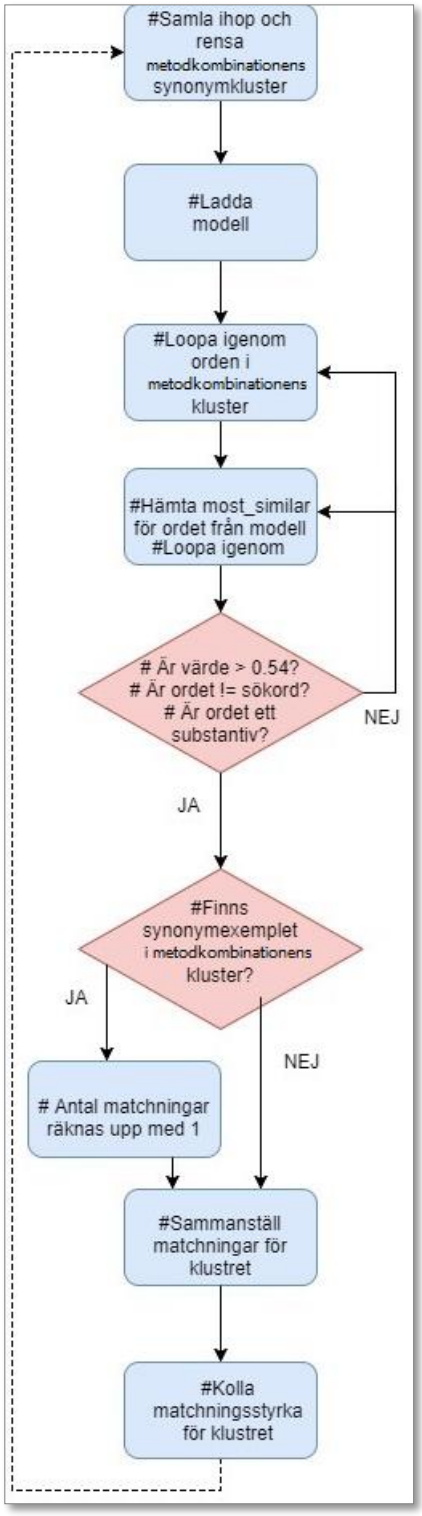
3.2.3 Testning.

Testningen utfördes i syfte att besvara frågeställningen: Kan en Word2vec-modell erbjuda ett större täckningsområde på så vis att modellen får fram andra synonymer än metodkombinationen? Denna fråga besvarades genom att jämföra modellens synonymförslag med metodkombinationens. Då ingen guldstandard fanns tillgänglig så kunde resultatet av testningen endast visa hur lika eller olika förslagen var och inte vilka förslag som var mest korrekta. Om förslagen skulle visa sig vara olika skulle detta kunna indikera att metoderna tillsammans skulle kunna uppnå en större räckvidd än den ena enskilt och om förslagen skulle visa sig vara lika indikerar detta att metoderna i kombination inte ger en större räckvidd.

För att jämföra metodkombinationens synonymer med de framtagna av modellen så användes Python 3.7. Först togs synonymer fram till vart och ett av orden i testdatamängden, det vill säga metodkombinationens kluster. Figur 3 illustrerar en simpel beskrivning av processen. Ett sökord (ett ord från synonymklustret) gav upphov till ett kluster bestående av ordets synonymförslag. Metodkombinationens kluster bearbetades innan på så vis att alla bokstäver gjordes till gemener för att inte göra skillnad på samma ord på grund av gemener eller versaler, därefter sorterades också dubbletter av ord i ett och samma kluster bort för att undvika att ett och samma ord poänggavs mer än en gång. För att ta fram synonymförslag ur modellens datamängd användes cosinuslikheten. De synonymförslag från modellen som inte var substantiv sorterades bort eftersom metodkombinationens kluster endast skulle innehålla substantiv. Resterande synonymförslag framtagna av modellen för ett specifikt ord jämfördes sedan med synonymförslagen som metodkombinationen tagit fram för samma ord.

Jämförelsen skedde utifrån hur väl synonymförslagen matchade varandra, alltså hur lika de föreslagna synonymklustren för ett visst ord var. Modellens synonymförslag för ett specifikt ord itererades och om samtliga av modellens förslag för detta ord utgjorde metodkombinationens synonymförslag så ansågs det vara en matchning på 100%. Detta gällde även om modellens förslag innehöll ytterligare ord utöver de som fanns i metodkombinationens kluster, det viktiga var alltså om modellen lyckades hitta de synonymer som metodkombinationen hittat oberoende av om modellen hittat flera synonymförslag. Detta gjordes för att hantera de fall då modellens förslag innehöll flera ord än metodkombinationens kluster. Om antalet matchningar dividerats med summan av matchningar och icke matchningar så hade resultatet berott av antalet synonymförslag från modellen men eftersom syftet var att jämföra modellens förslag med metodkombinationens så ansågs möjliga matchningar istället bero av antalet ord i metodkombinationens förslag. Därav beräknades resultatet genom att dividera antalet matchningar för ett kluster med antalet ord i metodkombinationens synonymkluster minus det sökta ordet.

För att hantera flerordstermer så togs mellanslag bort och byttes mot ett bindestreck för att ordet skulle kunna hanteras av modellen, ordet hanterades därefter som ett vanligt ord.



Figur 3. Illustration av flödet i koden. Detta flöde itereras en gång för varje synonymkluster från metodkombinationen till dess att alla kluster itererats. Modellens förslag itereras i en inre loop en gång per sökord.

3.2.4 Utvärdering.

En utvärdering utfördes manuellt av Fodina i syfte att avgöra om modellens synonymförslag var relevanta eller inte vid de tillfällen som modellens förslag inte matchade metodkombinationens förslag samt för att på så vis besvara frågeställningen: *Kan en word2vec-modell användas som en metodik för metodkombinationens synonymidentifiering?*

Denna frågeställning besvarades genom att utvärdera synonymförslagen från metodkombinationen och modellen för att konstatera vilka synonymförslag, då ingen matchning mellan klusterna fanns, som var mest korrekta. En indikation på att modellen skulle kunna användas som en metodik för metodkombinationens synonymidentifiering skulle då visas genom att modellen erhöll många korrekta då metodkombinationen istället erhöll många felaktiga vilket motiverar detta metodval. Om metodkombinationen erhöll många korrekta och modellen många felaktiga så hade det istället visat det motsatta. Frågeställningen skulle kunna besvarats genom att endast utvärdera modellens synonymförslag men eftersom synonymförslagen för utvärdering valts ut utifrån en 0% matchning så blev det även av vikt att se vilken av modellen och metodkombinationen som fick fram flest relevanta förslag, därav användes metodkombinationens förslag som en baslinje i detta fall. Utan denna jämförelse hade osäkerheten kvarstått om metodkombinationens kapacitet i förhållande till modellen.

För att öka trovärdigheten i modellens förslag så användes endast de synonymförslag som påvisade en cosinuslikhet på minst 0,54. Dessa ansågs vara godtyckliga förslag baserat på de mänskliga uppskattningarna som presenterades i *The WordSimilarity-353 Test Collection.* Värdet togs fram genom att alla ord som innehöll minst en uppskattning på 10 valdes ut och ett medelvärde av den lägst skattade siffran i dessa mängder beräknades och resulterade i ett värde på 5,272 för testsamling1 och 5,526 för testsamling2. Detta gav ett medelvärde på 5,399 vilket i procent motsvarar ca 54 %. Den procentuella skalan gav då en cosinuslikhet på 0,54.

Utvärderingen utfördes genom att slumpmässigt välja ut ett antal sökord (eng. target word). Dessa valdes från den klustermängd där inga matchningar fanns. För att vidare säkerställa att lika många synonymförslag från metodkombinationen och modellen utvärderades så valdes 250 slumpade synonymförslag från modellen samt 250 slumpade synonymförslag från metodkombinationen för att utvärderas. Detta val gjordes dels för en jämn datafördelning och också för att minska tidskomplexiteten för utvärderingen. Att utvärdera ett kluster åt gången kan i vissa fall, beroende på varierad klusterstorlek, vara mycket tidskrävande. Därav gjordes valet att utföra utvärderingen ordvis istället för klustervis.

Utvärderingsmaterialet skapades i ett Excel-ark innehållande fem olika kolumner. Den första kolumnen innehöll ett unikt ID. Sökordet presenterades i den andra kolumnen och ett av dess synonymförslag (eng. Synonym Suggestion) presenterades i den tredje kolumnen. Den fjärde kolumnen var tom vid utskick och skulle innehålla poäng ifyllda under utvärderingens gång. Poängskalan löpte mellan -1 till 1 där 1 poäng tilldelades korrekta synonymförslag, -1 felaktiga och vid tvivel tilldelades istället 0 poäng. Tvivel klassades som ord som i vissa sammanhang kan användas som synonymer men egentligen per definition inte är synonyma. Detta kan exempelvis utgöras av typer av homonymer, hyperonymer och hyponymer. Även om utvärderaren inte kunde avgöra om förslaget var rätt eller fel på grund av exempelvis felaktiga ord så kunde denne välja att se förslaget som tvetydigt och därmed poängsätta detta ord till 0. Den femte och sista kolumnen var inte synlig för utvärderarna då denna innehöll källan för synonymförslaget, alltså ”Metodkombination” eller ”Modell” beroende på varifrån förslaget kom. Detta gjordes för att undvika att utvärderarna medvetet eller undermedvetet skulle påverkas av att veta varifrån synonymförslaget kom.

Excel-arket skickades ut till Fodina, för illustration av utvärderingens formatering vid utskick se Figur 4 nedan. En och samma person fattade beslut om poäng och fyllde i arket. Samma utvärdering utfördes av tre olika utvärderare i syfte att öka trovärdigheten och säkerheten i bedömningarna eftersom synonymer kan vara svårbedömda. Med anledning av detta utfördes även en analys av den interna reliabiliteten hos utvärderarna genom Fleiss Kappa (Fleiss, 1971).

Utvardering_Fodina.xlsx

	A	B	C	D	E	F
1	ID	Target word	Synonym suggestion	Score		
2	7	ssl/tls	gnupg			
3	68	zoology	ichthyology			
4	671	mismanagement	appendage			
5	1623	annex	outgrowth			
6	1855	audacity	chutzpah			
7	1965	saddle	cantle			
8	2127	malware	rootkit			
9	2134	ctp	pyrophosphorylase			
10	2270	banteng	chevrotain			
11	2308	sctp	ospf			
12	2434	snowfall	rainfall			

Figur 4. Utvärderingen som den såg ut vid utskick. Kolumn A innehöll ID, kolumn B innehöll sökordet (eng. Target word), kolumn C innehöll ett synonymförslag till det specifika sökordet och kolumn D innehöll den tomma kolumnen där poängen skulle fyllas i. I kolumn E placerades källan efter utvärderingen för att jämföra metodkombinationen och modellens förslag.

4. Resultat

I detta kapitel presenteras resultaten av samtliga deluppgifter vilka utgörs av validering, testning och utvärdering.

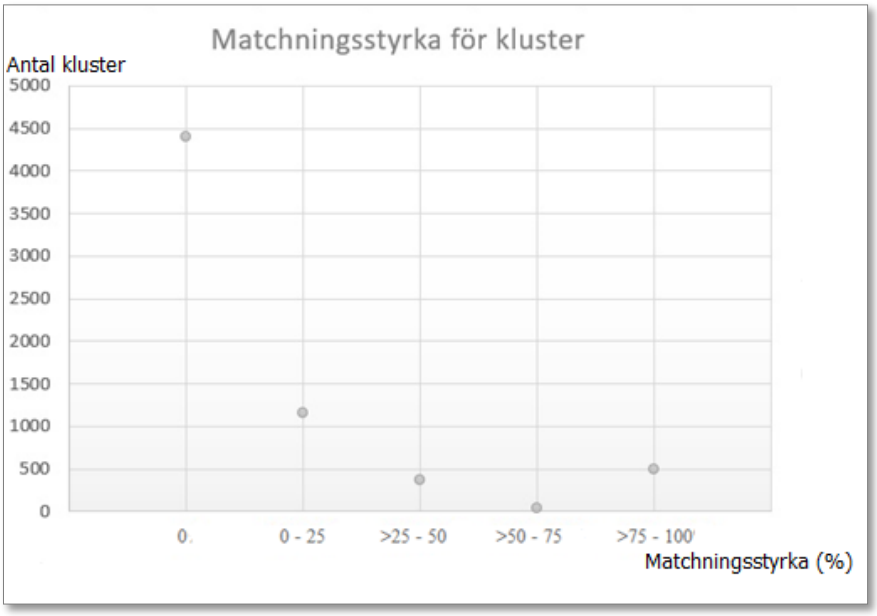
4.1 Validering

Resultatet av valideringen visade på att modellen skattade korrekt inom det mänskliga spannet i 59,86 % av fallen då testsamling1 användes och 60,73 % av fallen då testsamling2 användes, M = 60,30 %. I testsamling1 var det 11 ordpar som inte hittades av modellen och som därmed exkluderades från resultatet. I testsamling2 exkluderades 9 ordpar av samma anledning.

Vid utförande av Spearmans korrelationer mellan de mänskliga skattningarna och det beräknade värdet för cosinuslikheten från modellen påvisades en positiv korrelation både för Testsamling1 och Testsamling2. För Testsamling1 visades en signifikant mycket stark positiv korrelation, r_s = 0,70 och $p < 0,001$. Testsamling2 visade på en något svagare men signifikant stark korrelation med r_s = 0,65 och $p < 0,001$.

4.2 Testning

Resultatet av testningen visade på att 6 462 ord, vilket motsvarar 14,59 % av den totala ordmängden, bearbetades och därmed gav upphov till lika många kluster. Resultatet visade även på att 2 068 ord, motsvarande 32,00 % av den totala bearbetade ordmängden (N = 6 462) hade matchande synonymförslag från modellen och från metodkombinationen. 497 synonymkluster vilket motsvarade 7,96 % matchade helt eller nästan helt, med en matchningsstyrka på mer än 75 % och mindre eller lika med 100 %. Slutligen var det 4394 kluster vilket motsvarade 68,00 % av den totala klustersamlingen som inte matchade alls. I Figur 5 sammanställs resultatet av matchningsstyrka för samtliga kluster. Vid testningen var det 44 298 ord som inte fanns i modellens vokabulär och därmed har dessa ord uteslutits från resultatet.

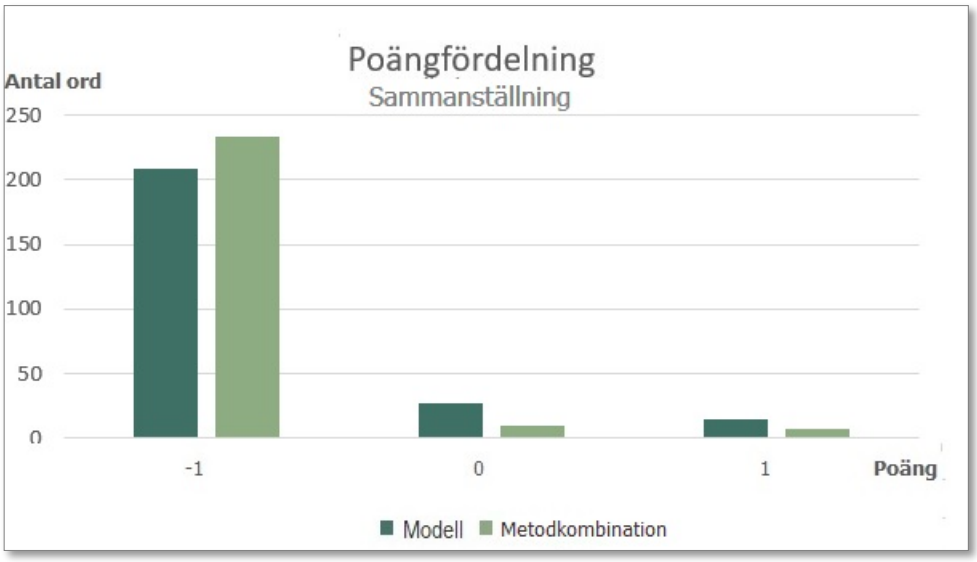


Figur 5. Fördelning av matchningsstyrka för kluster där x-axeln representerar matchningsstyrkan i procent och y-axeln antalet kluster. 4394 kluster (68,00 %) matchade inte alls, 1155 kluster (17,87 %) matchade till max 25 %, 372 kluster (5,76 %) matchade till max 50 %, 44 kluster (0,68 %) matchade till max 75 % och slutligen 494 kluster (7,69 %) matchade upp till 100 %.

4.3 Utvärdering

Den manuella utvärderingen visade att modellens synonymförslag (*N* = 250) innehöll fler korrekta förslag (*M* = 14,33) överlag med 5,73 % mot metodkombinationen (*N* = 250) (*M* = 7,67) med 3,07 %, i de kluster där inga matchningar mellan dem fanns. Antalet tvetydiga ord förekom i större utsträckning i modellens förslag (*M* = 27,33) gentemot metodkombinationens förslag (*M* = 9,33). Antalet felaktiga synonymförslag var därmed flest i metodkombinationens förslag (*M* = 233) mot modellens (*M* = 208). Figur 6 nedan illustrerar poängfördelningen baserat på medelvärde från samtliga utvärderingar. Den totala poängen baserat på medelvärden från samtliga tre utvärderingar summerades till -193,67 för modellen och -225,33 för metodkombinationen.

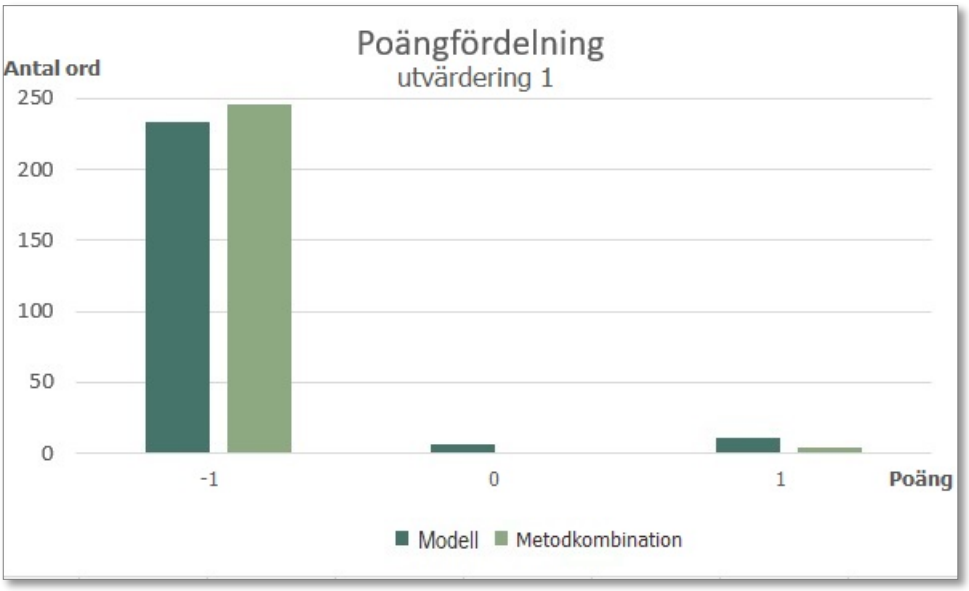
Fleiss Kappa visade på en befintlig men mycket svag enighet (eng. slight agreement) bland utvärderarna, Kappa = 0,19, 95 % CI (0,06, 0,33).



Figur 6. Sammanställning av poängfördelningen där x-axeln representerar de olika poängalternativen för modellen respektive metodkombinationen och y-axeln representerar antalet ord. Figuren är baserad på medelvärdet från samtliga utvärderingarna.

4.3.1 Utvärdering 1.

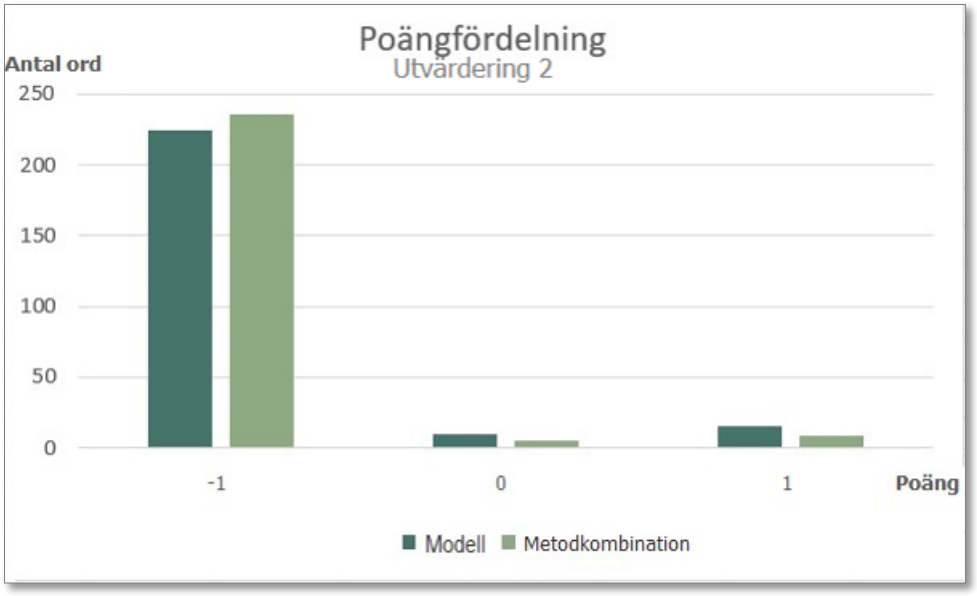
Den första utvärderingen visade att 4,4 % av modellens förslag var korrekta och 1,6 % av metodkombinationens. Av det totala antalet synonymförslag från modellen (N = 250) så ansågs 11 förslag vara korrekta, 233 ansågs vara felaktiga och 6 ansågs vara tvetydiga. Av det totala antalet synonymförslag från metodkombinationen (N = 250) ansågs 4 synonymförslag vara korrekta, 246 ansågs vara felaktiga och inga tvetydigheter konstaterades bland metodkombinationens synonymförslag. Figur 7 nedan illustrerar poängfördelningen från utvärdering 1.



Figur 7. Poängfördelning för utvärdering 1 för modellen och metodkombinationen där x-axeln representerar poängalternativen för modellen respektive metodkombinationen och y-axeln antal ord.

4.3.2 Utvärdering 2.

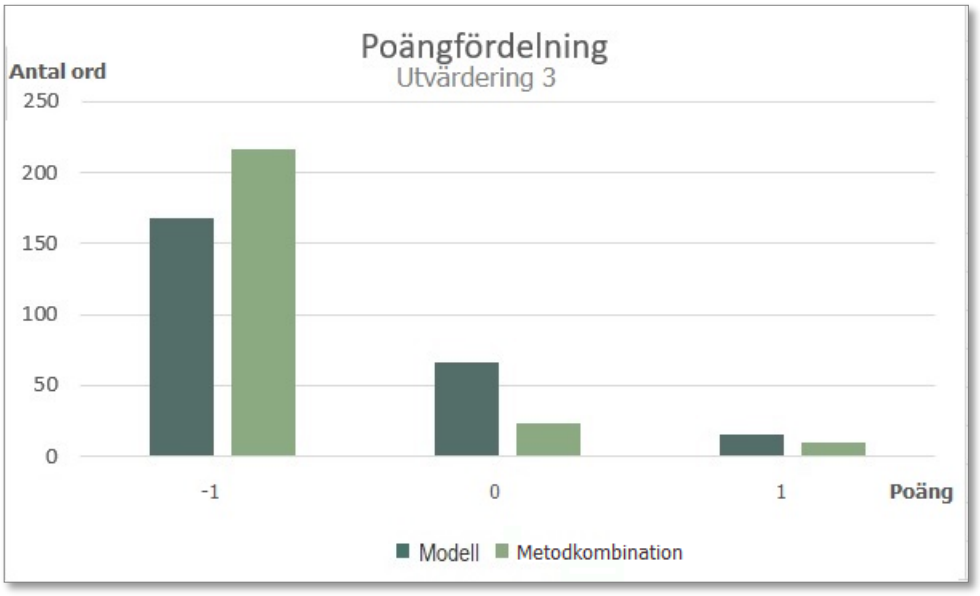
Den andra utvärderingen visade att 6,4 % av modellens förslag var korrekta och 3,6 % av metodkombinationens. Av det totala antalet synonymförslag från modellen (*N* = 250) ansågs 16 synonymförslag vara korrekta, 224 ansågs vara felaktiga och 10 ansågs vara tvetydiga. Av det totala antalet synonymförslag från metodkombinationen (*N* = 250) ansågs 9 förslag vara korrekta, 236 ansågs vara felaktiga och 5 ansågs vara tvetydiga. Figur 8 nedan illustrerar poängfördelningen från utvärdering 2.



Figur 8. Poängfördelning för utvärdering 2 för modellen och metodkombinationen där x-axeln representerar poängalternativen för modellen respektive metodkombinationen och y-axeln antal ord.

4.3.3 Utvärdering 3.

Den tredje utvärderingen visade att 6,4 % av modellens förslag var korrekta och 4,0 % av metodkombinationens. Av det totala antalet synonymförslag från modellen (*N* = 250) ansågs 16 synonymförslag vara korrekta, 168 ansågs vara felaktiga och 66 ansågs vara tvetydiga. Av det totala antalet synonymförslag från metodkombinationen (*N* = 250) ansågs 10 förslag vara korrekta, 217 ansågs vara felaktiga och 23 ansågs vara tvetydiga. Figur 9 nedan illustrerar poängfördelningen från utvärdering 3.



Figur 9. Poängfördelning för utvärdering 3 för modellen och metodkombinationen där x-axeln representerar poängalternativen för modellen respektive metodkombinationen och y-axeln antal ord.

5. Diskussion

I detta kapitel kommer resultatet att diskuteras både övergripande och i detalj inkluderat felkällor, metodval och förslag på vidare studier inom området.

5.1 Resultattolkning

Den övergripande tolkningen av samtliga resultat innehåller en förhållandevis stor osäkerhet. För att vidare avgöra vad resultatet egentligen innebär för de individuella uppgifterna så diskuteras och tolkas validering, testning och utvärdering separat.

5.1.2 Validering.

Resultatet av valideringen visade på tydliga kopplingar mellan modellens skattningar och mänskliga skattningar vilket vidare tyder på att denna Word2vec-modell faktiskt kan användas för att ta fram rimliga synonymförslag vilket var frågeställningen. Det kan dock diskuteras varför resultatet inte blev högre, det vill säga varför modellens värde inte överensstämde med de mänskliga skattningarna i en ännu större utsträckning. Detta kan dock bero på att skattningarna är baserade på likheter mellan ord och när människor utsätts för denna typ av uppgifter så tas flera språkliga aspekter hänsyn till. Valideringsdatamängden innehöll bland annat de mänskliga skattningarna av orden ”*king*” och ”*queen*” där det fanns människor som skattat dessa som 10. De ansåg alltså att orden var likvärdiga, vilket de ur flera aspekter är, men de är inte synonyma vilket är vad modellens förhoppning är att hitta. Därav kan det argumenteras för att resultatet som uppnåddes var tillräckligt högt eftersom ett högre värde skulle kunna innebära att modellen inte alls framtog synonymer utan snarare ord som liknar varandra i andra avseenden. Ett lågt värde på resultatet hade dock kunnat innebära att modellen inte alls fick fram rimliga skattningar vilket troligen skulle medföra att den heller inte var användbar för att ta fram synonymer. Ytterligare en anledning med valideringen var att kunna generalisera resultatet från testningen trots att modellen tränats på den data som metodkombinationen använt för att extrahera sina synonymkluster, därav var träningsdata och testdata nära relaterade. Denna validering kan då stärka att modellen är lämpad för annan data än den som användes vid testningen.

5.1.3 Testning.

Resultatet av testningen visade på att modellens synonymförslag överlag skilde sig från metodkombinationens vilket därmed innebär att modellen och metodkombinationen tillsammans skulle kunna utgöra ett större täckningsområde eftersom modellen får fram andra synonymförslag än metodkombinationen vilket var frågeställningen. Resultatet som fås fram kan, förutom faktiska skillnader i synonymförslagen, även bero på skillnaden i antalet förslag per ord, det vill säga klusterstorlek, eftersom detta kunde variera. För vissa ord blev antalet förslag från metodkombinationen fler än från modellen och möjligen tvärtom. Vidare innebär detta att måttet på olikheterna mellan synonymförslagen även innefattar olikheter orsakade på grund av färre eller flera ord per synonymkluster, vilket också är en olikhet och därmed kan det argumenteras för att denna skillnad faktiskt ska räknas med precis som den gjort. Med testningens utformning erhålls alltså en lägre matchningsstyrka om modellen hittat färre synonymförslag än de som finns i metodkombinationens kluster men om det är tvärtom så kan matchningsstyrkan fortfarande uppnå 100%. Denna utformning valdes eftersom metodkombinationens synonymkluster användes som en baslinje för modellen. Med tanke på att förhoppningen var att öka täckningsområdet så hade det varit positivt om modellen hittat fler synonymförslag än metodkombinationen men om samtliga av metodkombinationens förslag också fanns i modellens så skulle detta framgå. Detta visades genom en matchningsstyrka på 100% i dessa fall men hade även kunnat kompletterats med en räknare för vardera kluster för att vidare avgöra vilken som innehöll flest synonymförslag.

Många ord fanns inte i modellens vokabulär vilket medförde att stora datamängder räknades bort. Vidare kan detta indikera att orden som faktiskt fanns kanske inte tillhörde de mest vanliga orden. Därmed finns en risk att ordvektorerna tillhörande orden kan vara baserade på väldigt få meningar vilket medför en osäkerhet vid framtagning av synonymer. En word2vec-modell kräver, som tidigare nämnt, en stor datamängd att träna på för att prestera till sin fulla potential vilket bland annat har att göra med att ju mer data varje vektor baseras på desto mer troligt är det att den faktiskt representerar verkligheten. Det kan argumenteras för att parametern *min_count* vid träning därav bör vara så hög som möjligt för att säkerställa att representationen av verkligheten blir så korrekt som möjligt dock vill man även ta till vara på så mycket data som möjligt. Att sätta *min_count* till ett högre värde innebär även att ord som inte uppfyller detta minimumkrav helt räknas bort och senare hamnar inom kategoriseringen OOV-ord. OOV-ord är också något man vill undvika och därav blir det en avvägning vid parametersättningarna.

Det kan dock argumenteras för att osäkra vektorer är bättre än OOV-ord och därmed hade en sänkning av parameterinställningen *min_count* varit att föredra. Att öka denna hade troligen gett upphov till ännu fler OOV-ord vilket inte är önskvärt. Eventuellt kunde någon form av justering i modellen ha implementerats i syfte att hantera OOV-ord men då detta ofta ger upphov till slumpade värden på ordvektorerna så utfördes inte detta. En annan möjlig orsak till det höga antalet OOV-ord kan också vara med anledning av att ingen rensning utöver dubbletter utfördes för metodkombinationens kluster. Detta kan vidare ha medfört att några av sökorden faktiskt bestod av felaktiga ord som därav rättvisligen saknade vektorrepresentationer i modellen samt synonymförslag. Trots det höga antalet OOV-ord så ansågs den datamängd som fanns i modellens vokabulär och därmed bearbetades ändå godtycklig trots att majoriteten av den totala datamängden faktiskt fick räknas bort.

5.1.4 Utvärdering.

Resultatet av den manuella utvärderingen visade på att modellen överlag presterade marginellt bättre än metodkombinationen vid synonymidentifieringen då ingen matchning mellan de olika klustren fanns. Dock med den lilla andelen rätt som modellen påvisade i sina synonymförslag så finns ingen stark indikation på att modellen skulle kunna användas som en metodik för metodkombinationens synonymidentifiering vilket var frågeställningen. Osäkerheten stärks ytterligare av den mycket svaga enigheten mellan utvärderarna vilket vidare kan indikera att resultatet är involverat av slumpen och därmed inte generaliserbart. Resultatet visade även på att både metodkombinationen och modellen erhåller många felaktiga synonymförslag då metoderna är oeniga. Modellen erhöll förvisso en bättre summerad poäng än metodkombinationen men resultatet av den totala poängfördelningen berodde också starkt på att modellen erhöll betydligt fler tvetydiga synonymförslag. För modellens del kan antalet felaktiga och tvetydiga synonymförslag bero på att förslagen inte är kontrollerade på annat sätt än att cosinuslikheten ska vara större än 0,54. Det kan alltså vara så att de korrekta förslagen finns med i förslagsmängden men att de hamnar i skymundan på grund av det höga antalet felaktiga förslag. Med tanke på att valideringen ändå indikerade att modellen faktiskt kan få fram rimliga förslag så stärks teorin om att de rätta förslagen kan finnas med i synonymförslagen. Det kan dock också vara så att synonymförslag framtagna i testningen är tillhörande ord som är mycket mer ovanliga än de som användes vid valideringen. Återigen kan detta orsaka problem med ordvektorer som är baserade på väldigt få meningar och därmed i sin tur leverera felaktiga synonymförslag på grund av att ordvektorerna inte överensstämmer med

den verkliga representationen. Detta skulle även kunna orsaka att synonymförslagen för orden i testningen innehåller mer felaktiga bedömningar än orden från valideringen. Det är oavsett tydligt att både synonymkluster från modellen och metodkombinationen kräver en bearbetning för att rensa bort så kallade orimliga förslag.

5.2 Metodval

Många beslut handlar om avvägningar där fördelar och nackdelar måste balanseras. Ofta tvingas det fattas beslut där en fördel måste väljas bort för att göra utrymme åt en annan. Några av dessa avvägningar och beslut diskuteras och motiveras nedan vilket inkluderar val av datamängd och Word2vec-modell.

5.2.1 Datamängd.

Vid val av datamängd hade det bästa självklart varit att både använt en stor datamängd men också att den var domänspecifik. Då detta inte var en möjlighet tvingades denna avvägning att utföras. För att i så stor utsträckning som möjligt kunna utesluta problematikerna med antonymer så valdes en stor datamängd som inte var domänspecifik trots att detta talar emot resultatet av en del studier som menar på att domänspecificitet är viktigare än just en stor datamängd. Detta val gjordes även för att underlätta validering och utvärdering av modellen och säkerställa att en befintlig metod möjliggjordes för just effektiv utvärdering. Med domänspecificitet som krav för datamängden begränsas mängden tillgänglig data markant och det blir betydligt svårare att få tag på. Detta gäller både en tillräckligt stor datamängd men också testmängder som innehåller ord som finns i modellens vokabulär. Baserat på ovanstående så valdes en stor datamängd då detta reducerade flera eventuella problem som i sin tur troligen hade påverkat modellen negativt.

5.2.2 Word2vec-modell.

Vid val av Word2vec-modell stod det främst mellan en egentränad eller en färdigtränad. Den största fördelen med en egentränad modell är att parameterinställningar kan justeras fritt utifrån önskemål och behov vilket utgör större kontrollmöjligheter. Risken med att istället välja en färdigtränad modell är att detaljer kring träningen inte kan erhållas i samma utsträckning. För att avgöra användbarheten av en Word2vec-modell inom synonymidentifiering så är dock valet av egentränad eller färdigtränad inte av så stor betydelse. En Word2vec-modell är oavsett parameterinställningar tillräcklig för att uppnå studiens syfte. Det viktiga var att testa en

word2vec-modell som en första utvärderingsform och mer precisa parameterinställningar kan då tas i beaktning vid ett senare skede för eventuella förbättringar. Dock så kan parameterinställningar vid träning självklart ha inverkan på resultatet och i syfte att öka studiens validitet så valdes en modell vars parameterinställningar låg innanför det rekommenderade spannet i så hög utsträckning som möjligt. Risken att någon parameter enskilt skulle påverka minimerades på så vis.

Det är också tidskrävande att träna en egen modell på stora datamängder vilket var ytterligare en anledning till att en färdigtränad modell valdes och därmed kunde tid istället prioriteras att lägga på andra delar av denna studie som för syftet var mer relevanta.

5.3 Felkällor

Det finns många faktorer som kan ha bidragit till att resultatet påverkats. Dessa faktorer varierar inom de olika deluppgifterna som utgörs av träning, validering, testning och utvärdering och kommer därav diskuteras individuellt.

5.3.1 Träning.

För syftet med denna studie kan de flesta parameterinställningarna anses ha varit inom önskvärt spann men osäkerheten finns alltid vilket utgör parameterinställningarna till en möjlig felkälla. *Min_count,* som använde det förinställda värdet på 5 hade eventuellt valts till ett lägre värde om justeringsmöjligheten fanns, främst för att minska den stora mängden OOV-ord. Men för det övergripande resultatet hade detta troligen inte haft någon positiv inverkan utan hade troligen medfört en ännu större osäkerhet för representativiteten av de befintliga vektorerna. En annan parameter som eventuellt hade justerats till ett lägre värde var parametern *size* som under träningen var satt till ett värde på 300 vilket kan anses vara lite högt utifrån det rekommenderade spannet eftersom det valda värdet då ligger precis i överkant. Ett för högt värde kan utgöra en stor risk för överanpassning och det kan därmed vara fördelaktigt att hålla denna parameter så låg som möjligt. Enligt Tezgider, Yıldız, och Aydın (2018) så har parametern *size* dock visat sig ha en relativt liten inverkan på övergripande resultat och därav bör inte det något höga värdet ha en stor betydelse i detta fall heller eftersom värdet ändå låg inom det rekommenderade spannet.

5.3.2 Validering.

En möjlig felkälla vid valideringen är att det fanns ord som exkluderades vilket kan påverka resultatet. Dock var den exkluderade ordmängden liten i förhållande till den bearbetade

mängden och bör därför inte ha någon större betydelse för resultatet av valideringen av modellen. Orsaken till att totalt 20 ordpar exkluderades kan bero på att dessa var så pass olika att det inte gick att beräkna dess likheter med hjälp av cosinuslikheten. Detta eftersom olikheter teoretisk sett ska vara utplacerade på positiva respektive negativa planet av vektorrymden och cosinuslikheten används främst i det positiva planet.

5.3.3 Testning.

För att på bästa sätt besvara om en Word2vec-modell kan erbjuda ett större täckningsområde på så vis att modellen får fram andra synonymförslag än metodkombinationen hade testningen eventuellt kunnat göras annorlunda. Som tidigare nämndes hade det varit av intresse att kontrollera antalet synonymförslag från modellen respektive metodkombinationen för att vidare avgöra vilken som uppnådde störst täckningsområde. Det kan dock argumenteras för att antalet synonymförslag inte var relevant eftersom inga beslut om rätt och fel kunde göras i det läget. Ett högt antal synonymförslag utgör inte en större räckvidd om samtliga också är felaktiga. Därav ansågs den utförda testningen, innehållande jämförelsen mellan kluster, vara tillräcklig.

En annan viktig aspekt av testningen är att jämförelsen genomfördes på olika typer av kluster. Modellens synonymkluster skapades genom parvisa relationer på så vis att ett sökord hela tiden utgjorde utgångspunkten och synonymförslag togs fram utifrån detta sökord. De kluster som var framtagna av metodkombinationen bestod istället av sammanslagningar av flera olika parvisa relationer på så vis att ett synonymförslag kunde ge upphov till ytterligare synonymförslag som därigenom inte utgick från något sökord. På så vis bör jämförelsen visa på stora skillnader precis som den gjorde men detta kan alltså bland annat förklaras av dessa olika klustertyper.

5.3.4 Utvärdering.

Den manuella utvärderingen erhöll ett mycket osäkert resultat. Detta kan till viss del bero på komplexiteten inom mänsklig synonymidentifiering vilket medförde att de olika personerna bakom utvärderingen angett olika poäng för en stor mängd av synonymförslagen. Detta visar på en stor osäkerhet i vad som faktiskt är riktigt vilket även stärktes med den låga enigheten beskriven av Fleiss Kappa. Utvärderingar utförda av människor utgör alltid en risk men i detta läge då det handlar om synonymidentifiering så fanns inga andra alternativ. Det finns ingen tydlig guldstandard eftersom många ord är så pass tvetydiga att man utan kontext

inte kan säkert svara på vilka semantiska relationer som är riktiga. Detta kan även förklara den höga oenigheten hos utvärderarna.

På grund av den lilla datamängd som valdes ut till den manuella utvärderingen i förhållande till den totala datamängden så finns det en stor risk att resultatet av den manuella utvärderingen är missvisande. Resultatet stärks dock eftersom samtliga ord var framtagna med hjälp av slumpen vilket ökar generaliserbarheten men optimalt hade varit en större datamängd vilket inte var möjligt på grund av tidskomplexitet.

Det kan diskuteras om det var fördelaktigt eller inte att datamängden utgjordes av endast ord tillhörande kluster där inga matchningar fanns. Det är troligt att de synonymkluster som uppvisade en matchningsstyrka i de högre spannen innehåller fler korrekta alternativ då detta faktiskt innebär att metodkombinationen och modellen föreslagit samma ord. Att därmed slumpa ord från den mängden hade troligen gjort att ett bättre resultat erhållits. Men eftersom hela poängen med studien var att ta fram en metod som hittade andra synonymförslag än metodkombinationen blev det mer intressant att jämföra de kluster som skilde sig mest från varandra.

Viktigt att poängtera är också att utvärderarna ombads behandla synonymförslagen som om det var faktiska förslag vilket de inte var. Detta eftersom ingen rensning utöver minimumvärdet på 0,54 för modellens förslag och den nödvändiga anpassningen av metodkombinationens utfördes. Därav utgjordes datamängderna, framförallt från metodkombinationen av många förslag som troligen rensats bort av Termograph i ett senare skede. Detta val av utvärderingsform var dock nödvändigt för att minska subjektiviteten i bedömningarna så mycket som möjligt genom att endast ange rätt, fel eller tvetydigt. Större valmöjligheter hade troligen gett upphov till en ännu större osäkerhet och oenighet för utvärderarna.

5.3.5 Allmänt.

För synonymidentifiering med vektorrymdsmodeller kan ord och dess olika komplexitetsnivåer utgöra en felkälla. Synonymer är visserligen utbytbara med varandra i meningar men beroende på vilken typ av mening det är så väljs gärna ord som passar in i meningen utifrån svårighetsgraden i språket. Synonymer uppfyller ofta olika syften och kan därmed skilja sig enbart i just komplexitet vilket kan göra det svårare att upptäcka synonymer med hjälp av vektorrymdsmodeller. Om flera ord i en mening skiljer sig i komplexitet från en annars liknande mening så kan detta medföra att eventuella synonymer i dessa meningar inte hittas. Det kan vara två meningar som säger samma sak men den ena är skriven i exempelvis

talspråk och den andra akademiskt och därmed verkar de helt olika för en vektorrymdsmodell som baserar vektorerna utifrån liknande meningar.

5.4 Vidare studier

För att ta vid och utveckla denna studie samt ytterligare undersöka användningsområdet för Word2vec-modeller inom synonymidentifiering kan flera tillvägagångssätt övervägas. Bland annat kan konstateras att det finns exempel på synonymförslag som förvisso inte är synonymer men som ändå har semantiska egenskaper. Detta är något som starkt vill undvikas inom synonymidentifiering men som förekommer även inom mänsklig synonymidentifiering och uppfattning och är särskilt vanligt hos människor med afasi. Där har man sett tydliga kopplingar mellan hur hjärnan har bearbetat ord som är semantisk relaterade på liknande sätt. För att undvika detta vid träning av en modell bör först icke-synonyma men nära besläktade ord semantiskt, så som antonymer, hyperonymer och hyponymer, hanteras. Denna hantering utgör dock en problematik av minst lika stor grad av komplexitet som synonymidentifiering. Det är därför troligt att en metod ensam inte är tillräcklig för att hantera en komplexitet på denna nivå. Vidare studier för synonymidentifiering kommer därav att innebära att ta fram metoder för vardera delområde inom semantiska relationer såsom hantering av antonymer, hyperonymer och hyponymer.

Vidare kommer även den befintliga problematiken med homonymer att kräva en lösning vilket därmed utgör ytterligare en komplex studie. För samtliga av dessa utmaningar kan Word2vec-modeller användas och tränas i det specifika syftet för att därefter eventuellt kombineras. Dock blir då en risk att varje specifik modell även innefattar de övriga problematikerna och att en modell inte kan lösa dessa. Detta eftersom problematikerna går i varandra och överlappar vilket skapar en så kallad ond cirkel. Därav blir det svårt att avgöra vilken av dessa komplexa problem som eventuellt bör hanteras först och prioriteras före de övriga.

Utifrån denna enorma komplexitet som synonymidentifiering medför så kan dock ett första steg vara att träna modellerna på mer vanligt förekommande ord eftersom dessa då bör ha vektorer baserade på en tillräckligt stor datamängd. Valideringen i denna studie tyder på att Word2vec-modeller kan utgöra en lösning på synonymidentifiering specifikt för vanligt förkommande ord och det skulle därför vara betydelsefullt att ta detta vidare och även kombinera med de andra problematikerna som ett första försök att skapa en kombinerad metod.

Intressant vore också att undersöka vad en kombination av modellen och metodkombinationen skulle få fram efter en eventuell rensning av orimliga synonymförslag

med hjälp av Termograph. Detta eftersom skillnaden mellan deras kluster var så påtaglig och utan att ha noggrant jämfört kan ingen uttala sig om vad kombinationen kan komma att åstadkomma.

Med hänsyn till resultaten av testningen så kan eventuellt en kombination av de båda metoderna användas för att rensa synonymklustren på så kallade orimliga förslag. Att förslagen behöver rensas visades av resultaten av testningen och utvärderingen tillsammans eftersom majoriteten av förslagen inte matchade och vidare var det många felaktiga förslag som erhölls just vid de tillfällen då ingen matchning mellan klustren fanns. Därav skulle denna matchningsprocess kunna användas i syfte att rensa bort så kallade orimliga förslag vid de tillfällen då matchningsstyrkan mellan kluster är 0% eller då enskilda synonymförslag inte överlappar.

6. Slutsats

Det kan konstateras att synonymidentifiering sannerligen är en komplex uppgift och det krävs mer än denna tränade word2vec-modell för att lyckas prestera tillräckligt bra för att utgöra ett komplett verktyg inom synonymidentifiering för komplexa termer i dagsläget. Det som kan konstateras är att modellen enligt valideringen har förutsättningar att prestera och därav kan de något svaga resultaten av testningen och utvärderingen möjligen bero av utomstående faktorer.

Resultaten av denna studie är dock lite för osäkra för att användas som stöd för att vidare uttala sig om framtida möjligheter med Word2vec inom synonymidentifiering. Word2vec-modellen får förvisso fram rimliga synonymförslag och kan eventuellt användas tillsammans med metodkombinationen för ett större täckningsområde vid enstaka tillfällen i enlighet med resultaten. Men med hänsyn till de höga antalet fel som påvisades i utvärderingen så tyder detta på att oenigheten mellan modellen och metodkombinationen innebär en stor osäkerhet utan en tydlig vinnare. Vidare kan detta resultat nyttjas för framtida arbete, exempelvis för att rensa kluster då metodkombinationen och modellen inte matchar.

Med hänsyn till de problematiker som ytterligare komplicerar synonymidentifieringen, såsom hyponymer, hyperonymer och homonymer samt i med denna studies resultat i beaktande så krävs en metod som, utöver synonymidentifiering, även lyckas hantera de tillkommande problematikerna. Troligtvis krävs det ytterligare specifika metoder som hanterar dessa problem för att kunna komplettera de brister som indikeras både i metodkombinationens kluster och i de kluster framtagna av Word2vec-modellen.

7. Referenslista

Andersson, L., Grennberg, A., Hedberg, T., Näslund, R., Persson, L., Sydow, B. & Söderkvist, I. (1999). *Linjär Algebra med Geometri*. (2 ed.). Lund: Studentlitteratur AB.

Berardi, G., Esuli, A. & Marcheggiani, D. (2015). Word Embeddings Go to Italy: A Comparison of Models and Training Datasets. *IIR - 6th Italian Information Retrieval Workshop (Cagliari, Italy, 25-26 May 2015).*

Dridi, A., Gaber, M., Azad, R. & Bhogal, J. (2018). k-NN Embedding Stability for word2vec Hyper-Parametrisation in Scientific Text. *21st International Conference*, DS 2018, Limassol, Cyprus, October 29–31, 2018, Proceedings. 10.1007/978-3-030-01771-2_21.

Fares, M., Kutuzov, A., Oepen, S. & Velldal, E. (2017). Word vectors, reuse, and replicability: Towards a community repository of large-text resources. *Proceedings of the 21st Nordic Conference on Computational Linguistics*. 271-276.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G. & Ruppin, E. (2001). Placing Search in Context: The Concept Revisited. *ACM Transactions on Information Systems, 20(1):116-131*.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological bulletin, 76(5), 378.

Ganesan, K. (2018). How to get started with Word2vec – and then how to make it work. Hämtat 2019-03-06 från https://medium.freecodecamp.org/how-to-get-started-with-word2vec-and-then-how-to-make-it-work-d0a2fca9dad3?fbclid=IwAR3axPIdTZlZVdSAWUfMGSOIHx3J-tidzyx6occCxduAuWI409pJcFaNLs0

Lai, S., Liu, K., He, S. & Zhao, J. (2016). How to Generate a Good Word Embedding. *IEEE Intelligent Systems, 31*(6). 5-14. doi: 10.1109/MIS.2016.45

Lee, S. & Lee, K. (2018). Design of the Korean Medicine Symptom Diagnosis System Using Word2vec. *2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT).* doi: 10.1109/CAIPT.2017.8320727.

Lei, K., Si, S., Wen, D. & Shen, Y. (2017). An enhanced computational feature selection method for medical synonym identification via bilingualism and multi-corpus training. 909-914. 10.1109/*ICBDA*.2017.8078771.

McHugh, M., L. (2012). Interrater Reliability: The Kappa Statistic. *Biochem Med, 22*(3), 276-282. PMID 23092060.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781v3 [cs.CL]*. https://arxiv.org/pdf/1301.3781.pdf

Milne, D. & Witten, I. H. (2008). An Effective, Low-cost Measure of Semantic Relatedness Obtained from Wikipedia Links. *AAAI PRESS*. 25-30. http://www.aaai.org/Papers/Workshops/2008/WS-08-15/WS08-15-005.pdf

Nguyen, A. K., Schulte im Walde, S. & Vu, T. N. (2017). Distinguishing Antonyms and Synonyms in a Pattern-based Neural Network. *arXiv:1701.02962v1 [cs.CL]*, https://arxiv.org/pdf/1701.02962.pdf

Nordic Language Processing Laboratory. (u.å). Models. Hämtat 2019-03-26 från http://vectors.nlpl.eu/explore/embeddings/en/models/

Nurifan, F., Sarno, R. & Wahyuni, S. C. (2018). Developing Corpora using Word2vec and Wikipedia for Word Sense Disambiguation. *Indonesian Journal of Electrical Engineering and Computer Science 12(3)*, 1239-1246, doi: 10.11591/ijeecs.v12.i3

Rong, X. (2014). word2vec Parameter Learning Explained. *arXiv:1411.2738v4 [cs.CL]*, https://arxiv.org/abs/1411.2738

Russell, S., Norvig, P. (2016). Artificial Intelligence: A Modern Approach (3 ed). Essex: Pearson Education Limited.

Saeed, I. J. (2015). *Semantics*. (4. Ed.). West Sussex: Wiley-Blackwell.

Sahlgren, M. & Lenci, A. (2016). The Effects of Data Size and Frequency Range on Distributional Semantic Models. *Conference on Empirical Methods in Natural Language Processing*, 975–980. doi: 10.18653/v1/D16-1099.

Scheible, S., Schulte im Walde, S. & Springorum, S. (2013). Uncovering distributional differences between synonyms and antonyms in a word space model. *Asian Federation of Natural Language Processing.* 489-497. http://aclweb.org/anthology/I13-1056

Strube, M. & Ponzetto, S. P. (2006). Wiki-relate! Computing Semantic Relatedness using Wikipedia. *AAAI'06 proceedings of the 21st national conference on Artificial intelligence, (2),* 1419-1424. http://new.aaai.org/Papers/AAAI/2006/AAAI06-223.pdf

Tezgider, M., Yıldız, B. & Aydın, G. (2018). Improving Word Representation by Tuning Word2vec Parameters with Deep Learning Model. *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*. doi: 10.1109/IDAP.2018.8620919

Wikimedia Foundations. (2017). Hämtat den 2019-04-10 från https://archive.org/details/enwiki-20171201.

Zhang, L., Li, J. & Wang, C. (2017). Automatic Synonym Extraction Using Word2vec and Spectral Clustering. *2017 36th Chinese Control Conference (CCC)*. doi: 10.23919/ChiCC.2017.8028251

Zhou, Z., Fu, B., Qiu, H., Zhang, Y. & Liu, X. (2017). Modeling medical texts for distributed representations based on Skip-Gram model. *2017 3rd International Conference on Information Management (ICIM)*. doi: 10.1109/INFOMAN.2017.7950392