



Universitatea Transilvania din Braşov
Facultatea de Matematică şi Informatică
Specializarea Tehnologii Moderne
în Ingineria Sistemelor Soft

LUCRARE DE DISERTAȚIE

Autor: Muşat Bogdan-Adrian
Coordonator: Conf. dr. Sasu Lucian-Mircea

Braşov
2017

Cuprins

1	Introducere	2
1.1	Motivația alegerii temei	4
1.2	Structura lucrării	4
1.3	Recunoaștere	4
2	Lucrări similare	5
2.1	Modele de regăsire	5
2.1.1	Duolingo	6
2.1.2	Facebook Messenger	6
2.2	Modele generative	7
3	Arhitectura	8
3.1	Rețele neurale artificiale	8
3.2	Rețele neurale artificiale recurente	9
3.3	Modele Sequence-to-Sequence	11
3.4	Hierarchical Recurrent Encoder-Decoder	12
3.4.1	Celula GRU	13
3.4.2	HRED bidirecțional	14
3.5	Word Embeddings	15
3.5.1	Word2vec	15
3.5.2	Word embeddings preantrenate	16
3.6	Predicție	16
3.6.1	Greedy	17
3.6.2	Beam search	17
4	Experimente și rezultate	19
4.1	Arhitectura 1.0	19
4.2	Arhitectura 1.5	20
4.3	Arhitectura 2.0	20
4.4	Arhitectura 3.0	20
4.5	Arhitectura 3.5	21

5	Posibile dezvoltări	23
5.1	Învățare prin întărire	23
5.2	Mecanisme de atenție	25
5.3	Setul de date	26
6	Concluzii	28

Capitolul 1

Introducere

Procesarea de limbaj natural a reprezentat dintotdeauna una dintre marile provocări ale comunităților științifice de pretutindeni, scopul final fiind realizarea unui sistem suficient de inteligent încât să fie capabil să comunice precum omul. În acest scop, în anul 1950 a fost lansat testul Turing care presupune o conversație om-mașină și un interlocutor care participă la această discuție, acesta nefiind capabil să distingă omul de mașină. Acest test încă stă în picioare după aproape 70 de ani, niciun sistem fiind capabil încă să-l doboare.

Recent, diverse companii de succes precum Google, Facebook, Microsoft, Apple și alții au realizat un demers semnificativ spre dezvoltarea unor sisteme inteligente de comunicație numite chatbots. La ora actuală, cei mai mulți sunt orientați către suport tehnic, scopul lor fiind tocmai de a imita sprijinul pe care o persoană reală îl poate oferi unui client. Un sistem computațional inteligent poate fi folosit pentru a răspunde multor cereri simultane, iar costurile de întreținere sunt mici – practic, odată ce sistemul este dezvoltat, este necesară doar expunerea lui, de exemplu ca serviciu web. Cu cât tehnologia și munca de cercetare în această direcție progresează, cu atât sistemele de genul acesta devin din ce în ce mai inteligente și mai apropiate de ceea ce un om este capabil să ofere. Precum automatizarea industrială de producere a vehiculelor reprezintă un standard în era contemporană, asistenții conversaționali vor deveni un standard în anii ce vor urma.

Majoritatea modelelor actuale se bazează pe oferirea unor răspunsuri predefinite. Aceste tipuri de sisteme pot returna doar răspunsuri existente, nefiind capabile de a genera text nou. Pe de altă parte, avem sistemele bazate pe inteligență artificială generative care, precum omul, pot emite răspunsuri bazate pe experiențe anterioare. Vorbim deci de evoluția unor modele de

chatbots bazate pe pattern matching către unele bazate pe modele generative. În ultimii ani, ramura inteligenței artificiale numită Machine Learning (învățare automată) a luat amploare în această direcție prin curentul numit Deep Learning. Acest curent a produs o mulțime de rezultate spectaculoase atât în direcția procesării naturale de limbaj cât și a procesării de imagini.

Rolul principal al unui chatbot îl reprezintă capacitatea de a “înțelege” informația primită de la o persoană pentru a produce un raspuns cât mai plauzibil. Însă cum procesează un calculator o limbă? Pentru a răspunde la această întrebare, se va face o analogie la cum învață un om o limbă. Pornește de la anumite cuvinte de bază iar apoi pe baza acestora, învață cuvinte tot mai complexe. Începe să creeze fraze prin care leagă aceste cuvinte precum și gramatica specifică limbajului. Practic, totul se bazează pe o anumită experiență cumulativă. Conversația devine astfel o modalitate ce facilitează și impulsionează deprinderea limbajului: omul este pus în fața unor contexte de utilizare, fenomen ce întărește deprinderea de utilizare a cuvintelor sau expresiilor individuale. Fiecare cuvânt nou învățat reprezintă o experiență către învățarea în continuare a limbajului. Încă un aspect reprezentativ uman este capacitatea de a întreține o conversație pe termen lung cu un alt participant în intermediul unui context, lucru dificil de capturat pentru un sistem artificial.

Abordările de chatbot actuale își propun imitarea acestei modalități de învățare pentru un sistem computațional. Modelarea curentă cea mai eficientă pentru o astfel de problemă este oferită de către Deep Learning. Folosind arhitecturi de rețele neurale artificiale mult mai complexe decât cele shallow de la începutul anilor 2000 și plecând de la seturi de instruire masive - în cazul de față, corpusuri de text cu cât mai multe fraze în limba dorită se pot crea modele generative care pot produce continuarea unor propoziții, fraze etc. Ideea din spatele abordării Deep Learning este că sistemul devine mai performant pe măsură ce volumul de date crește. Pentru a procesa o asemenea cantitate de date de instruire este nevoie de putere computațională pe măsură. Asta presupune pe scurt, capacitate hardware. Milioanele de calcule efectuate pentru modelarea unui limbaj de către sistem nu permit folosirea unui procesor, chiar multicore de ultimă generație, deoarece este considerat a duce la sugrumarea procesului de instruire.

În locul microprocesoarelor se preferă folosirea plăcilor grafice (GPU). Inițial, acestea au fost dezvoltate pentru rulare rapidă de jocuri, dar potențialul lor a fost rapid intuit și exploatat prin programare paralelă. Deoarece placa

video conține mult mai multe nuclee de procesare (câteva mii, comparate cu cele 4-8 nuclee tradiționale dintr-un microprocesor actual), este preferată programarea și rularea modelelor computaționale pe GPU. În ultimii ani au dezvoltat o multitudine de biblioteci care facilitează unui programator dezvoltarea de aplicații de Machine Learning pe GPU: Tensorflow, Theano, Caffe, Keras etc.

1.1 Motivația alegerii temei

Testul Turing reprezintă o provocare pe care mulți cercetători din acest domeniu o abordează constant cu noi idei tot mai inovative, având din ce în ce mai mult succes în ultimii ani. Chiar și la nivelul unui model generativ de jucărie, acest subiect reprezintă unul dintre cele mai fierbinți topicuri la ora actuală pentru comunitatea Deep Learning. Acest fapt, precum și cercetarea ideilor state of the art folosite în sistemele actuale reprezintă factorul motivațional principal în alegerea temei de cercetare.

1.2 Structura lucrării

TODO

1.3 Recunoaștere

Mulțumiri Universității Transilvania din Brașov care a finanțat acest proiect în scopul achiziționării de hardware necesar pentru realizarea acestuia.

Capitolul 2

Lucrări similare

În anul 1950, matematicianul Alan Turing a propus un test care pune la încercare abilitatea unei mașini computaționale de a manifesta inteligență echivalentă cu cea umană. Un evaluator uman judecă o conversație între un om și o mașină desemnată să genereze răspunsuri cât mai naturale. Evaluatorul este conștient că unul dintre partenerii angrenați în discuție este o mașină. Conversația este limitată doar la text astfel încât rezultatul să nu depindă de abilitatea calculatorului de a genera sunete. Dacă evaluatorul nu poate diferenția mașina de om, putem spune că aceasta a trecut testul. Până în prezent, acest test rămâne încă în picioare.

Există două mari abordări la ora actuală de a genera limbaj cât mai natural. Avem în primul rând modele de tip regăsire, unde răspunsul este stocat într-o sursă de date și returnat pe baza unor metode de pattern matching. Cea de-a doua variantă o reprezintă modelele generative, care produc un răspuns dinamic folosind diverse metode din teoria probabilităților.

2.1 Modele de regăsire

Modelele de acest tip folosesc o sursă de date care conține numeroase răspunsuri predefinite. Răspunsul este ales folosind o metodă euristică pentru o potrivire cât mai bună cu putință, luând în considerare intrarea și contextul. Tipul de euristică folosit poate fi ceva simplu precum o expresie bazată pe o regulă de potrivire sau ceva mai complex cum ar fi un clasificator de Machine Learning. Se poate deduce foarte ușor că aceste sisteme nu generează text nou. O problemă uriașă a acestor modele o reprezintă incapacitatea de a reacționa la cazuri nemaiîntalnite pentru care nu există un răspuns potrivit. Acestea au totuși avantajele lor. Datorită sursei de date cu răspunsuri create

de oameni, aceste metode nu produc erori gramaticale. În prezent, acestea reprezintă abordarea sigură pentru problemele în care răspunsul este unul sensibil, într-un domeniu precum cel medical, de exemplu.

2.1.1 Duolingo

Populara aplicație de învățare a limbilor străine Duolingo (Figura 2.1) folosește o abordare interesantă cu privire la chatbots. Aceasta dorește să își ajute utilizatorii să practice o nouă limbă prin conversații cu chatbots. Având în vedere că o conversație este considerată a fi printre cele mai bune moduri de a învăța o limbă străină, utilizatorii Duolingo pot vorbi cu bot-ul oricât de mult își doresc, iar acesta îi va corecta și le va propune răspunsuri potrivite. Mai mult de atât, poate estima progresul utilizatorului pentru a-și crește nivelul de dificultate, păstrând astfel constantă provocarea.



Figura 2.1: Conversație cu chatbot folosind Duolingo

2.1.2 Facebook Messenger

Facebook s-a alăturat întru totul afacerii conversaționale, astfel încât și-a transformat aplicația Messenger într-un business de mesagerie. Compania a integrat plățile peer-to-peer în Messenger în anul 2015, apoi urmând să lanseze un API pentru chatbots, astfel încât business-urile să poată crea interacțiuni pentru clienți. Poți comanda flori, să navighezi printre ultimele trend-uri în materie de modă, să comanzi Uber, toate dinăuntrul chat-ului de Messenger (Figura 2.2).

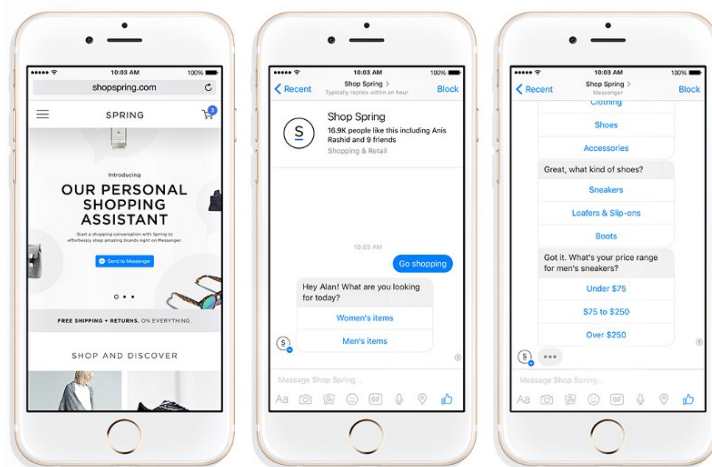


Figura 2.2: Cumpărături de haine folosind chatbot-ul companiei Spring pe Facebook Messenger

2.2 Modele generative

Spre deosebire de modelele anterioare, cele generative nu se bazează pe răspunsuri predefinite ci generează noi răspunsuri pornind de la zero. Modelele generative folosesc de obicei tehnici din Machine Translation, dar în loc de a traduce dintr-o limbă într-alta, vom ”traduce” de la o intrare la o ieșire (răspuns). Acestea oferă o mai bună impresie de comunicare cu un om real. Totuși, ele sunt extrem de greu de antrenat, sunt predispuse la erori gramaticale (în special unde lungimea propoziției este mai mare) și necesită o cantitate mare de date de antrenare.

Prezentul este încă sub semnul întrebării pentru acest tip de model, însă în următorii ani, acestea vor căpăta tot mai multă atenție și popularitate devenind tot mai performante. Dacă un chatbot va doborî testul Turing, șansele ca acesta să fie generativ sunt destul de mari. Deoarece modelele generative reprezintă o arie de cercetare încă nefinisată, acestea nu sunt folosite momentan în producție. În capitolul 3 vor fi prezentate diferite arhitecturi de rețele neurale, capabile să modeleze conversații.

Capitolul 3

Arhitectura

Modelarea unui limbaj reprezintă o sarcină extrem de dificilă pentru un calculator. Recente progrese în aria Deep Learning au făcut posibile diverse dezvoltări în această direcție, însă lucrurile sunt departe de a fi rezolvate. Arhitectura folosită pentru construcția chatbot-ului prezentat în această lucrare va fi expusă precum un bloc de construcții, plecând de la noțiunile de bază, până la arhitectura finală.

3.1 Rețele neurale artificiale

O rețea neurală artificială (RNA) reprezintă o paradigmă bazată pe procesare de informații a cărei inspirație provine din sistemul nervos biologic. Precum creierul uman, o RNA este compusă dintr-un număr mare de elemente de procesare interconectate (neuroni) lucrând la unison pentru a rezolva diverse probleme. RNA, precum oamenii, învață din exemple. Învățarea în sistemele biologice implică ajustarea conexiunilor sinaptice care există între neuroni. Acest principiu se aplică și acestor rețele.

Ca și modelare matematică propriu-zisă, o RNA poate fi observată în Figura 3.1. Avem o intrare reprezentată în figură de vectorul n -dimensional (x_1, x_2, x_3, x_4) . Această intrare reprezintă caracteristicile unui eșantion care face parte dintr-un set de date. Straturile intermediare se numesc straturi ascunse, iar rolul lor este să producă o abstractizare cât mai complexă a datelor de intrare, folosind diverse funcții cu activare neliniară. Ultimul strat se numește strat de ieșire, reprezentând eticheta (clasa) vectorului de intrare.

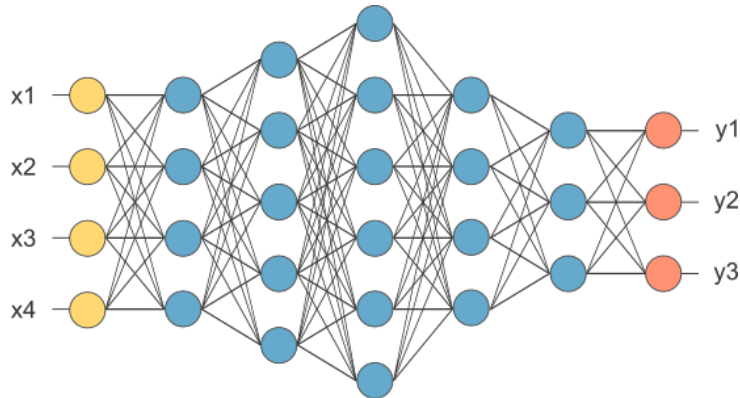


Figura 3.1: Exemplu de rețea neurală artificială

Ceea ce de fapt acest tip de rețele încearcă să învețe sunt ponderile dintre straturile sale. Se pleacă de la un set de ponderi alese aleator¹ și se ajustează folosind algoritmul de propagare înapoi a erorii până când rețeaua se stabilizează.

TO CONTINUE

3.2 Rețele neurale artificiale recurente

După cum se poate observa, o RNA este utilă atunci când intrarea este formată dintr-un singur element, de exemplu o imagine. Modelarea unui limbaj însă presupune ca intrările să fie fraze. O frază este formată din mai multe elemente (cuvinte), deci o RNA clasică nu poate modela o astfel de intrare. Soluția pentru această problemă este oferită de rețelele neurale artificiale recurente (RNAR). Acestea sunt folosite pentru a modela secvențe unde există o dependență temporală (fraze, serii de timp). Fiecare intrare curentă din secvență este condiționată de cele precedente. O astfel de rețea se poate observa în Figura 3.2. Straturile ascunse într-o RNAR au rolul de a păstra o captură a tot ceea ce s-a oferit ca intrare până în momentul curent. Valoarea fiecărui strat ascuns curent depinde astfel atât de intrarea curentă cât și de ieșirea stratului ascuns anterior. Putem considera toată această modelare ca pe o probabilitate condiționată $P(x_n|x_1, x_2, \dots, x_{n-1})$, unde în cazul unei fraze x_1, x_2, \dots, x_n reprezintă cuvintele acesteia.

¹Valorile ponderilor sunt de obicei subunitare iar inițializarea este făcută urmând diverse principii matematice bine definite. Pentru cazuri simpliste, inițializarea poate fi totuși făcută folosind o distribuție uniformă.

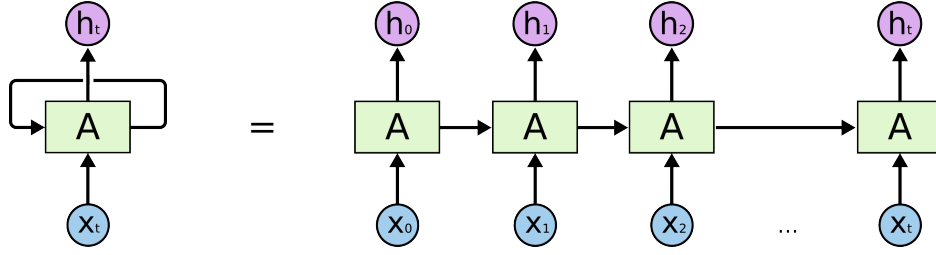


Figura 3.2: Exemplu de rețea neurală artificială recurentă

Algoritmul de reglare a ponderilor se numește propagarea înapoi în timp a erorii. O problemă serioasă care a apărut odată cu introducerea acestor rețele se numește anularea gradientilor². În cazul în care secvența de intrare are o lungime mare, informația nu reușește să se propage în timp deoarece gradientul ponderilor tinde să devină 0. Mecanismul care înlătură această problemă se numește Long Short Term Memory (LSTM). Principiul este ca prin folosirea unor porți de transmitere a informației, gradientul să fie stabilizat. Acest mecanism a reprezentat un avans spectaculos pentru Deep Learning, permițând modelarea secvențială cu reținere a informației pe o perioadă îndelungată de timp. Ecuațiile 3.1 descriu calculele pentru porțile LSTM.

$$\begin{aligned}
 f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\
 i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\
 o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\
 c_t &= f_t \circ c_{t-1} + i_t \circ \tanh(W_c x_t + U_c h_{t-1} + b_c) \\
 h_t &= o_t \circ \tanh(c_t)
 \end{aligned} \tag{3.1}$$

Poarta f_t se numește poartă de uitare (forget gate). Rolul ei este să stabilizească câtă informație se va uita din trecut. Poarta i_t este poarta de intrare (input gate). Aceasta delimitează care este cantitatea de informație care se păstrează din intrarea curentă. Poarta o_t este cea de ieșire (output gate). Ea controlează ce informație va fi transmisă către ieșire. c_t se numește starea internă a celulei LSTM. Aceasta este calculată ca o combinație între starea anterioară a celulei, poarta de intrare și cea de ieșire. h_t reprezintă ieșirea curentă a rețelei și depinde de poarta de ieșire și starea celulei LSTM.

²vanishing gradients

O RNAN aduce mai aproape ideea de modelare lingvistică, însă se poate observa că aceasta nu permite ca intrare decât o frază pe rând. Modelarea dorită în această lucrare este o pereche de forma întrebare-răspuns.

3.3 Modele Sequence-to-Sequence

Traducerea a reprezentat mereu un punct de interes în mediul procesării naturale de limbaj, constituind de altfel o provocare uriașă de-a lungul ultimilor ani. Până în anul 2014, majoritatea modelelor de traducere se bazau pe lanțuri Markov ascunse (Hidden Markov Models - HMM), totul urmând a se schimba odată cu introducerea unei noi arhitecturi în acel an de către Cho et Al. Noul tip de arhitectură se numea Sequence-to-Sequence (seq2seq) și urma să aducă îmbunătățiri spectaculoase atât pe partea de traducere cât și pe partea conversațională. Modelul este vizibil în Figura 3.3.

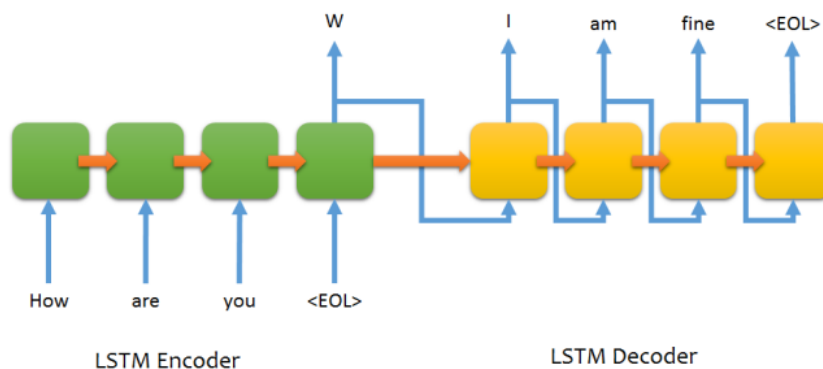


Figura 3.3: Modelul Sequence-to-Sequence

Aceast tip de arhitectură este împărțită în două jumătăți: codor (encoder) și decodor (decoder). Sarcina codorului este să primească o frază cu un număr variabil de cuvinte ca intrare iar unica sa ieșire ³ să fie o captură a întregii intrări (un vector de lungime fixă). Această captură poate fi privită ca o sumarizare a întregii fraze. Sarcina decodorului este de a învăța fraze pornind de la ieșirea codorului, care va deveni prima intrare din decodor, și restul intrărilor precedente. Intrarea curentă w_n în decodor este ieșirea de la timpul anterior, w_{n-1} . Modelarea se transformă astfel într-o probabilitate condiționată: $P(w_n|w_1, w_2, \dots, w_{n-1}, c)$, unde c reprezintă ieșirea codorului, deseori întâlnit sub numele de context în literatura de specialitate.

³Ieșirea ultimului element din secvență

Intrările în acest model sunt de forma întrebare-răspuns. Original folosit ca model de translație, unde intrarea pentru codor era fraza într-o limbă iar intrarea în decodor era fraza tradusă în limba dorită, acest principiu se poate aplica la fel de ușor pentru a modela o conversație. Spre deosebire de traducere unde totul se întâmpla punctual iar răspunsul nu depinde decât de intrarea curentă, o conversație este foarte dependentă de un context. Fără acest context, partenerul angrenat în discuție ar răspunde mereu luând în considerare doar ce i s-a spus în momentul de față, ignorând orice replică anterioară. Acesta nu este un comportament dorit în cadrul unei conversații și de aceea modelul seq2seq nu este destul de puternic de sine stătător pentru a modela o discuție.

3.4 Hierarchical Recurrent Encoder-Decoder

Modelarea contextului discuției reprezintă una dintre principalele nevoi în ceea ce privește o conversație care dorește să pară cât mai reală. Până recent, cea mai apropiată arhitectură care realiza acest lucru era modelul seq2seq, însă contextul era reținut doar la nivelul unei singure fraze. În anul 2016, Iulian Șerban et Al. a introdus rețeaua Hierarchical Recurrent Encoder-Decoder (HRED - Figura 3.4). Ea poate fi văzută ca o extensie peste seq2seq. Avantajul acesteia este că poate reține contextul discuției.

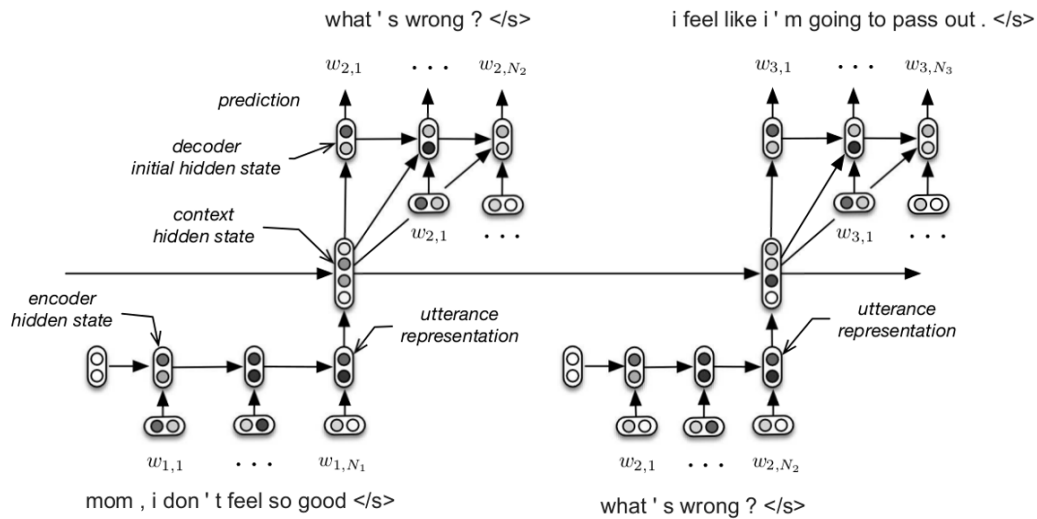


Figura 3.4: Hierarchical Recurrent Encoder-Decoder

După cum se poate observa, modelul este construit prin alăturarea mai multor rețele seq2seq legate între ele printr-o RNAR contextuală. După cum

sugerează și numele, rolul acestei RNAR este de a reține contextul discuției de-a lungul mai multor fraze. Spre deosebire de seq2seq, ieșirea codorului nu mai este oferită direct ca și primă intrare pentru decodor, aceasta trecând mai întâi prin RNAR contextuală. Pentru a fi mai ușor de înțeles ce se întâmplă, ne putem imagina toate cele n codoare și decodoare ca fiind simple unități de intrare, respectiv ieșire într-o RNAR obișnuită. Precum sunt într-o RNAR ieșirile dependente atât de intrarea curentă, cât și de cele precedente, prin analogie putem deduce că într-o rețea HRED, fiecare ieșire depinde de fraza curentă și de contextul discuției (frazele precedente).

3.4.1 Celula GRU

Gated Recurrent Unit sau GRU este un tip de celulă pentru o RNAR, menită să o înlocuiască pe cea LSTM. Ecuatiile din spatele celulei(3.2) sunt asemănătoare cu cele LSTM.

$$\begin{aligned} z &= \sigma(x_t U^z + s_{t-1} W^z) \\ r &= \sigma(x_t U^r + s_{t-1} W^r) \\ h &= \tanh(x_t U^h + (s_{t-1} \circ r) W^h) \\ s_t &= (1 - z) \circ h + z \circ s_{t-1} \end{aligned} \tag{3.2}$$

GRU are doar două porți: poarta r de resetare (reset gate) și poarta z de actualizare (update gate). Intuitiv, poarta de resetare determină cum se va combina noua intrare cu memoria precedentă iar cea de actualizare definește cât de multă memorie anterioară se păstrează. Principalele diferențe între GRU și LSTM sunt:

- GRU are doar două porți, pe când LSTM trei.
- GRU nu posedă o memorie internă c_t care să fie diferită de starea ascunsă expusă. Nu conține poarta de ieșire care este prezentă pentru LSTM.
- Poarta de intrare și cea de ieșire sunt cuplate de poarta de actualizare, iar poarta de resetare este aplicată direct stării ascunse anterioare. Astfel, responsabilitatea porții de resetare din LSTM este împărțită între cea de resetare și cea de actualizare de la GRU.
- Nu se aplică o a doua neliniaritate atunci când se calculează ieșirea.

Nu există o arhitectură câștigătoare clară între LSTM și GRU. Având mai puțini parametri (U și W), GRU se antrenează mai rapid și are nevoie

de mai puține date pentru a generaliza. Pe de altă parte, o cantitate mare de date exprimă o putere mai mare de modelare din partea LSTM și astfel se pot obține rezultate mai bune. Pentru realizarea acestei aplicații a fost aleasă o arhitectură de tip GRU.

3.4.2 HRED bidirecțional

Rolul codorului, precum a fost menționat mai devreme, este să captureze informația unei fraze într-un vector de lungime fixă. În orice limbaj, sensul unui cuvânt nu este stabilit doar de cuvintele precedente ci și de cele ce vor urma. Deoarece RNAR modelează cuvintele dintr-o frază pornind de la cuvântul w_1 la w_n , înțelesul din viitor al acestora este ignorat, iar captura frazei poate să nu fie îndeajuns de reprezentativă. Astfel, este propusă folosirea unei RNAR bidirecționale (Figura 3.5) pentru codor. Acest tip de rețea folosește două parcurgeri ale secvenței de intrare: cea înainte, care se desfășoară în mod obișnuit și cea inversă (de la w_n la w_1), pentru capturarea înțelesului din viitor al cuvintelor. Ieșirea pentru parcurgerea înainte va fi la timpul n iar cea pentru parcurgerea inversă va fi la timpul inițial. Vectorul de lungime fixă va fi format din concatenarea celor două ieșiri ale rețelei bidirecționale.

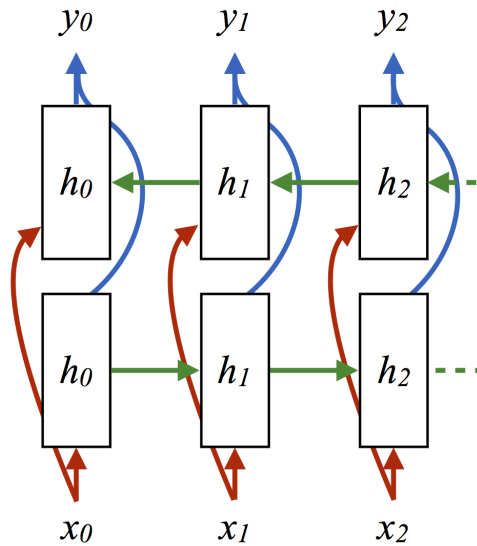


Figura 3.5: RNAR bidirecțională

3.5 Word Embeddings

Buna reprezentare a caracteristicilor unei intrări conduce la rezultate comparativ mai bune față de situațiile unde acestea sunt înfăptuite manual. Când vine vorba despre procesarea naturală de limbaj, buna reprezentare a cuvintelor din vocabular (word embeddings) înseamnă un factor decisiv pentru succesul final. O reprezentare evidentă ar fi folosirea codificării de tipul one-hot care presupune un vector de lungimea vocabularului, umplut cu elemente 0 iar poziția pe care se află cuvântul va fi marcată cu 1. Dacă lungimea vocabularului este mare, vectorul va deveni rar (sparse) și avem de a face cu noțiunea de blestemul dimensionalității⁴. Astfel, este necesară o codificare compactă, mai inteligentă, care să fie de asemenea semnificativă pentru semantica cuvintelor.

3.5.1 Word2vec

În anul 2013, Mikolov et al. a propus un model pentru word embeddings care se numește word2vec. În prezent, atunci când vine vorba despre modelare de limbaj natural, word2vec este soluția cea mai abordată în practică. Aceasta propune folosirea unor vectori de lungime fixă, compactă care totodată să captureze sensul cuvintelor. Ideea de bază este ca în spațiul n -dimensional folosit, cuvintele cu semantică asemănătoare să aibă un scor de similaritate cât mai mare. Ca și metrică de similaritate preferată pentru cuantificarea asemănării între reprezentarea a două cuvinte, câștigătoarea este de cele mai multe ori similaritatea cosinus (ecuația 3.3). Cel mai popular exemplu produs pentru a descrie puterea de modelare folosită de word2vec este *king – man + woman ≈ queen*. Figura 3.6 capturează ideea de word2vec vizualizat într-un spațiu bidimensional.

$$s(w_1, w_2) = \frac{\sum_{i=1}^n w_{1i}w_{2i}}{\sqrt{\sum_{i=1}^n w_{1i}^2}\sqrt{\sum_{i=1}^n w_{2i}^2}} \quad (3.3)$$

⁴Dimensionalitatea vectorului de intrare este extrem de mare iar arhitectura rețelei nu este capabilă să modeleze distribuția de probabilități a datelor

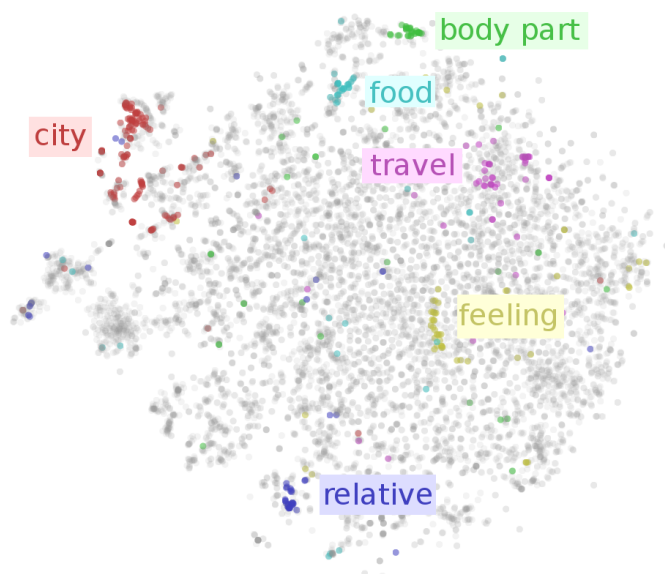


Figura 3.6: Word2vec vizualizat în spațiu 2D

3.5.2 Word embeddings preantrenate

În machine learning, de cele mai multe ori există arhitecturi deja antrenate pentru diverse sarcini precum clasificare de imagini, recunoaștere vocală etc. La fel este cazul și pentru word embeddings. Google a produs o antrenare pe un set de date care constă din miliarde de articole de știri, rezultând în urma acestuia word embeddings pentru aproximativ trei miliarde de cuvinte. Dacă se dorește antrenarea unei rețele care să învețe word embeddings pornind de la zero este posibil, însă sarcina este destul de dificilă iar rezultatul produs este dependent de natura cuvintelor din setul de date disponibil. Modelarea ideală ar fi ca aceste reprezentări ale cuvintelor să exprime cât mai bine realitatea înconjurătoare și nu un subset al acesteia. Din acest motiv, este preferată folosirea modelului preantrenat de word embeddings pusă la dispoziție de Google.

3.6 Predicție

Având modelul antrenat, acesta trebuie folosit pentru predicție, adică generarea unor răspunsuri adecvate pentru contextul discuției. Generarea cuvintelor este condiționată atât de context, cât și de cuvintele anterior generate. Spre deosebire de antrenare, în momentul predicției, secvența de ieșire este goală inițial (un vector de zerouri). Cuvintele sunt generate secvențial,

unul câte unul și adăugate pe poziția t la care a ajuns secvența. Deoarece stratul softmax al secvenței de ieșire returnează un vector de probabilități peste toate cuvintele din vocabular, pentru a returna cuvântul dorit se alege cel cu probabilitatea cea mai mare. Pornind de la această idee, există două variante de generare: metoda greedy și beam search.

3.6.1 Greedy

Precum în teoria clasică a algoritmicii, scopul metodelor greedy este să selecteze la fiecare pas optimul local. În cazul de față, optimul local reprezintă cuvântul cu probabilitatea de apariție cea mai mare. Prin această metodă nu este garantat la final că fraza produsă este cea mai bună cu putință, deoarece o altă combinație de cuvinte poate produce oricând o frază cu un scor mai bun⁵. Astfel, singurul avantaj al acestei metode este viteza de predicție rapidă. În practică, se preferă folosirea unor algoritmi mai complexi, capabili să aleagă fraza cea mai potrivită dintr-un set de fraze candidat.

3.6.2 Beam search

Algoritmul beam search folosește ideea de arbore de căutare pentru a genera fraze (Figura 3.6). În locul alegerii cuvântului cu probabilitatea cea mai mare la fiecare pas, se vor alege top k cuvinte cu cele mai mari probabilități. În literatură, k se mai numește și beam size. Fiecare nod (cuvânt) ales va produce un număr n_c de copii. În felul acesta este garantată generarea mai multor fraze candidat din care se va alege. Fiecare frază are un scor $S(f)$, care este de forma ecuației 3.3, unde $P(w_i)$ reprezintă probabilitatea cuvântului ales. O frază devine candidat atunci când cuvântul curent generat este simbolul de sfârșit de propoziție. În momentul în care algoritmul a terminat de generat toate frazele candidat, cea mai bună este determinată ca fiind cea cu scorul $S(f)$ cel mai mare.

$$S(f) = \frac{1}{n} \sum_{i=0}^n \log P(w_i) \quad (3.4)$$

Pentru ca acest arbore de căutare să nu capete o creștere exponențială, la fiecare pas se păstrează un număr egal cu beam size cele mai bune fraze

⁵De regulă se utilizează media aritmetică a sumei din logaritmul probabilităților cuvintelor frazei

iar restul sunt decartate. Acest algoritm nu asigură de departe optimul global, însă spațiul căutării oferit de folosirea unui arbore este mult mai bine dezvoltat și ales decât la metoda greedy. Cu cât beam size și numărul de copii generați sunt mai mari, cu atât algoritmul oferă soluții cât mai apropiate de optimul global. De asemenea, odată cu creșterea acestor parametri, crește și timpul de căutare și generare în arbore. De aceea, trebuie să existe un compromis între calitatea predicției și viteza de execuție a algoritmului.

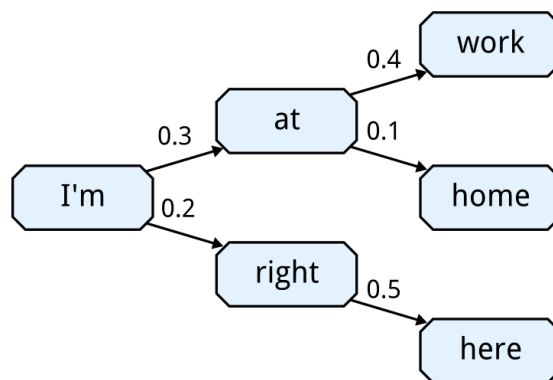


Figura 3.7: Arbore de căutare pentru algoritmul beam search

Capitolul 4

Experimente și rezultate

Acest capitol prezintă configurațiile arhitecturale încercate, cele care au avut succes precum și rezultatele finale obținute de către sistem. Pentru validarea modelului obținut s-a realizat o scalare progresivă a numărului de date de intrare, acesta fiind supus la două teste principale. Unul dintre teste a fost ca pentru puține exemple de intrare, modelul să fie capabil să învețe mot a mot să producă răspunsuri precum în setul de antrenare (overfitting). Un alt test a fost ca atunci când numărul de intrări este crescut considerabil, sistemul să nu producă întotdeauna răspunsuri monotone. Metrica de evaluare principală a sistemului a fost cea umană, ca și în cazul testului Turing.

4.1 Arhitectura 1.0

Arhitectura inițială presupunea folosirea unei dimensiuni maxime pentru o frază de 75 de cuvinte. Deoarece frazele au dimensiuni variabile de lungime, iar o RNAR acceptă dimensiuni fixe, pentru a completa fraza până la lungimea maximă aleasă, se umpleau spațiile lipsă cu vectori 0, procedeu denumit și padding în literatură. Fiecare frază se sfârșește cu un simbol special numit end token iar apoi urmează padding-ul. Fiecare strat de ieșire RNAR din HRED (codor, context și decodor) avea o dimensiune de 256 de neuroni. De remarcat este faptul că fiecare unitate structurală din HRED, rețelele recurente folosite pentru codor și decodor, foloseau ponderi shared¹. Stratul de embedding pentru cuvinte pornea cu învățarea de la 0, nefolosindu-se word embeddings provenite de la word2vec. Ieșirea din RNAR contextuală era concatenată cu reprezentarea fiecărui cuvânt din decodor. Ieșirea era mai apoi procesată de un strat ascuns cu activare tanh pentru expandarea

¹Fiecare codor din HRED folosea același set de ponderi, precum și în cazul decodurului

dimensiunii și rezultatul era concatenat cu restul secvenței de intrare pentru decodor. Algoritmul pentru învățarea ponderilor folosit a fost rmsprop. Pentru predicție s-a folosită varianta greedy descrisă în Capitolul 3. Această versiune performa bine atunci când erau date puține, însă odată cu creșterea volumului, modelul nu reușea să scaleze.

4.2 Arhitectura 1.5

Principala deosebire față de versiunea precedentă a fost folosirea word embeddings pentru reprezentarea cuvintelor și înghețarea stratului respectiv pentru a nu modifica ponderile preantrenate. S-a dovedit că această alegere a îmbunătățit atât timpul de antrenare, deoarece numărul parametrilor a fost redus considerabil, cât și performanța modelului, acesta fiind capabil de a captura un număr mai mare de date, dar totuși insuficient. Algoritmul pentru optimizarea ponderilor folosit a fost înlocuit cu adam folosind constanta de învățare implică (0.001), care se presupune că obține o convergență mai bună față de rmsprop. De asemenea s-a încercat folosirea simbolului special de început de frază denumit și start token, însă s-a observat că ieșirea contextuală era întotdeauna înclinată către acest simbol în timpul predicției, sistemul nemaifiind capabil să modeleze fraza dorită ca răspuns.

4.3 Arhitectura 2.0

Deoarece modelul eșua în a scala atunci când numărul de date creștea, presupunerea a fost că acesta nu era îndeajuns de complex pentru a reprezenta distribuția setului de antrenare. De aceea, s-a încercat folosirea unor straturi suplimentare alipite vertical² pentru codor și decodor deoarece teoria spune că de la strat la strat, modelul capturează reprezentări tot mai complexe ale intrării. Procesul de instruire a devenit astfel foarte lent, deoarece complexitatea arhitecturii a crescut, însă rezultatul final era nesatisfăcător.

4.4 Arhitectura 3.0

Arhitectura 3.0 a adus cu sine un progres important spre atingerea obiectivului final. S-a renunțat astfel la straturi stacked deoarece nu rezolvau problema scalabilității. Noua configurație pentru fiecare ieșire RNAR era 300 de neuroni pentru codor, 300 de neuroni pentru context și 300 pentru decodor.

²Stacked

S-a renunțat de asemenea la concatenarea ieșirii contextuale cu fiecare reprezentare a cuvintelor din decodor și trecerea acestora prin stratul ascuns cu activare tanh, deoarece nu mai era nevoie de expandarea dimensiunii. Astfel, ieșirea contextuală era folosită acum doar ca primă intrare T_1 pentru secvența decodurului, presupunerea fiind că influența contextuală se propagă oricum prin RNAR a decodurului. Încă un considerent pentru această decizie a fost faptul că ieșirea contextuală și reprezentarea word2vec a cuvintelor provin din distribuții diferite iar împreunarea lor nu reprezenta un factor corect din punct de vedere matematic. Desigur, mecanismele de atenție descrise în Capitolul 5 posedă o putere de modelare sporită față de simpla utilizare a contextului ca primă intrare pentru decodor. Încă o observație realizată care încetinea procesul de antrenare era folosirea unei fraze cu o lungime maximă de 75 de cuvinte. În urma analizei amănunțite a setului de date, s-a observat că majoritatea frazelor aveau o lungime scurtă. Astfel, s-a calculat valoarea de 80% (80 percentile) pentru lungimea maximă a frazelor și s-a determinat că aceasta este aproximativ 25. Sacrificiul realizat pentru a ignora cuvintele frazelor care depășesc lungimea de 25 de cuvinte a produs o scădere considerabilă a timpului de instruire. Odată cu schimbarea către această configurație, s-a adoptat algoritmul beam search de predicție. Deoarece spațiul de căutare devenea mai larg, răspunsurile au devenit și ele mai diversificate. Valorile pentru numărul de copii generați de fiecare nod și beam size au fost alese 2, respectiv 5. În teorie, cu cât aceste numere sunt mai mari, spațiul explorat crește și răspunsul devine mai calitativ. Însă creșterea considerabilă a acestor doi parametri aduce cu sine o viteză computațională lentă, de aceea trebuie făcut un compromis la nivelul acestor valori. În pofida acestor îmbunătățiri, sistemul încă nu scala în mod așteptat odată cu creșterea datelor.

4.5 Arhitectura 3.5

Deoarece setul de date conține subtitrări din filme, este un lucru obișnuit ca majoritatea răspunsurilor să fie monotone. Astfel, se creează o tendință pentru anumite cuvinte care apar foarte des, cele rare fiind aproape ignorate de către sistem. De aceea, sistemul eșuează în a furniza răspunsuri diversificate, care să conțină și cuvinte rare. Acest neajuns se poate rezolva prin folosirea ponderilor pentru cuvinte în momentul când este calculată valoarea funcției de eroare. Ponderea pentru fiecare cuvânt va fi numărul de apariții / numărul total de cuvinte. Prin această metodă se asigură că sistemul consideră influența egală a tuturor cuvintelor nemaiaivând o predispoziție către cele care apar des. Dimensiunea ieșirilor codor, context, decodor au fost alese ca fiind 500, 300, 500 pentru creșterea ușoară a complexității mo-

delului. În urma antrenării se putea observa că modelul rămâne blocat în anumite regiuni de minim local. Acest aspect a fost înlăturat prin reducerea progresivă a constantei de învățare de-a lungul instruirii, atunci când valoarea funcției de eroare nu scădea pe parcursul a cinci epoci consecutive cu mai mult de o valoare ϵ aleasă. Factorul de reducere a fost ales ca fiind 75%. Această combinație de ponderi împreună cu reducerea treptată a constantei de învățare au reprezentat succesul care au produs convergența modelului curent și totodată apariția rezultatelor satisfăcătoare.

Capitolul 5

Posibile dezvoltări

5.1 Învățare prin întărire

Învățarea prin întărire sau Reinforcement learning (RL) reprezintă o arie machine learning inspirată de psihologia comportamentală, a cărei principale preocupări este cum poate un agent software să îndeplinească anumite acțiuni într-un mediu înconjurător astfel încât să maximizeze o răsplată cumulativă. Tipul acesta de învățare diferă de metodele standard supervizate prin faptul că sistemului nu îi este prezentat niciodată perechi de forma intrare-ieșire, precum nici rezolvări ale unor acțiuni suboptimale. Focusul principal este asupra performanței on-line care presupune găsirea balanței ideale între explorare (dobândirea de cunoștințe noi) și exploatare (folosirea cunoștințelor curente).

Modelul de bază RL presupune:

- Un set de stări pentru mediu și agent S
- Un set de acțiuni A ale agentului
- O politică¹ de tranziție de la stări la acțiuni
- Reguli care determină răsplata scalară imediată a tranzițiilor
- Reguli care descriu observațiile agentului

¹policy

Un agent RL interacționează cu mediul în pași de timpi discreți. La fiecare pas t , agentul primește o observație o_t , care de obicei include o răsplată r_t . Apoi alege o acțiune a_t din setul de acțiuni posibile, care este trimisă mai departe către mediu. Mediul realizează o tranziție către starea nouă s_{t+1} , iar recompensa r_{t+1} asociată cu tranziția (s_t, a_t, s_{t+1}) este determinată (Figura 5.1).

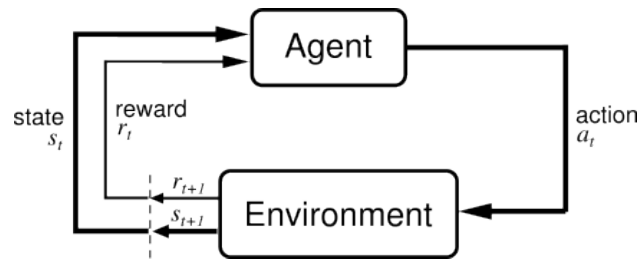


Figura 5.1: Tranziția stărilor în RL

Pentru a acționa optim, agentul trebuie să ia în considerare dependența pe termen lung a acțiunilor sale (să maximizeze recompensa viitoare), chiar dacă răsplata imediată este asociată cu o valoare negativă. Astfel, RL este potrivit pentru probleme unde există un compromis care include o recompensă între o dependență pe termen lung și una pe termen scurt. A fost aplicat cu succes în diverse situații precum controlul automat de roboți, telecomunicații, jocuri ATARI și recent jocul GO.

Jiwei Li et al. propun în articolul lor, Deep Reinforcement Learning for Dialogue Generation, utilizarea metodelor de RL pentru învățare de limbaj. Plecând de la aceeași arhitectură HRED, descrisă în capitolul 3, se poate ca prin folosirea unor reguli euristice de determinare a recompensei, pe baza răspunsurilor oferite de către chatbot, rețeaua să fie capabilă să ofere răspunsuri mai clare, contextuale și corect gramaticale. Astfel, articolul propune trei metode de recompensă pentru cuantificarea calității unui răspuns:

- r_1 , care recompensează modelul pentru răspunsuri care nu fac parte dintr-un subset de răspunsuri manual alese ca fiind monotone (I'm sorry, I don't know, Yes, No, etc.)
- r_2 , care recompensează sistemul pentru nerepetarea aceluiași răspuns consecutiv de mai multe ori la rând
- r_3 , care recompensează corectitudinea gramaticală și coerența răspunsului

Răsplata finală va fi o medie ponderată între cele trei recompense:

$$\lambda_1 r_1 + \lambda_2 r_2 + \lambda_3 r_3$$

unde $\lambda_1 + \lambda_2 + \lambda_3 = 1$, iar $\lambda_1 = 0.25, \lambda_2 = 0.25, \lambda_3 = 0.5$

Folosind RL, autorii respectivi demonstrează că rezultatele pot fi considerabil îmbunătățite. De aceea, extinderea modelului curent pentru a folosi RL reprezintă una dintre principalele posibile dezvoltări dorite.

5.2 Mecanisme de atenție

Mecanismele de atenție pentru o RNA sunt puternic bazate pe cele vizuale existente la om. Atenția vizuală umană, deși este bine studiată și există diverse modele, toate se rezumă la a fi capabile să se concentreze asupra unei anumite regiuni dintr-o imagine cu o rezoluție sporită, pe când imaginea de fundal este percepută cu o rezoluție mai scăzută, punctul focal fiind ajustat de-a lungul timpului.

Pentru a observa utilitatea acestora, se pleacă de la arhitectura standard seq2seq ca exemplu. Precum a fost descris în capitolul 3, codorul are rolul de a captura înțelesul întregii fraze de intrare într-un vector de lungime fixă, care mai apoi va fi folosit pentru inferență. Pare nerezonabil ca întreaga informație să fie stocată într-un singur vector, în special dacă lungimea frazei este mare. LSTM se presupune că rezolvă cu succes problema dependenței temporale a unei secvențe, însă în practică aceste neajunsuri încă apar atunci când lungimea acesteia este prea mare. O rezolvare plauzibilă poate fi oferită prin folosirea RNAR bidirecționale, care produc rezultate semnificativ mai bune.

Aici intervin mecanismele de atenție. Acestea nu mai capturează întreaga informație a codorului într-un singur vector fix ci permit decodorului să observe diferite părți ale frazei sursă la fiecare pas al generării. Ceea ce este important este că modelul învață singur ce informație trebuie selectată pe baza a ceea ce a fost produs până în acel moment și al intrării (Figura 5.2).

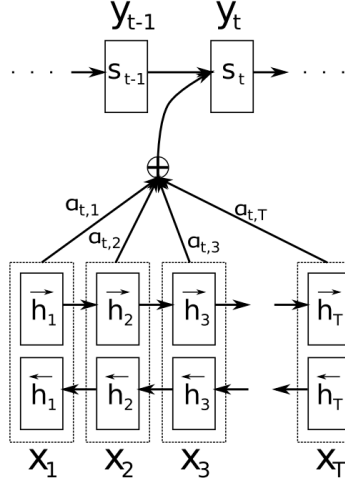


Figura 5.2: Mecanism de atenție

În figura de mai sus, y reprezintă ieșirea generată de către decodor iar x cuvintele corespunzătoare frazei de intrare. Aspectul cel mai important este că fiecare ieșire y_t depinde acum de o combinație ponderată (valorile a) a tuturor stărilor din codor, și nu doar de ultima. De exemplu, dacă $a_{3,2}$ este un număr mare, asta înseamnă că decodorul este foarte atent la cea de-a doua stare de intrare atunci când este produs al treilea cuvânt. Ponderile a sunt de obicei normalizate astfel încât suma lor să fie egală cu 1^2 .

Costul atașat acestor mecanisme este destul de mare. Pentru fiecare combinație intrare-ieșire trebuie calculată o valoare de atenție și stocată. Acest lucru este contraintuitiv, deoarece atenția umană se presupune că economisește resurse computaționale focalizând doar ce este important. Modelul prezentat analizează totul în detaliu înainte de a decide ce este important. Totuși, acest lucru nu a împiedicat ca mecanismele de atenție să devină atât de populare și să producă rezultate bune pentru o multitudine de sarcini.

5.3 Setul de date

Unul dintre aspectele esențiale de care puterea de procesare Deep Learning profită este cantitatea mare de date. Dacă modelele tradiționale de Machine Learning nu reușeau să exploateze creșterea volumului de date, modelele Deep Learning nu suferă de această neajuns (Figura 5.3). Acest fapt

²Distribuție a stărilor de intrare

este explicabil prin simplul fapt că modelele tradiționale nu erau suficient de complexe pentru a captura distribuția datelor de intrare. Datorită complexității sporite a modelelor deep learning, care au avut parte de o explozie de popularitate odată cu creșterea capacității computaționale, a memoriei pentru stocarea datelor și folosirea procesorului grafic pentru a paraleliza calculele pe mii de nuclee, acestea pot surprinde cu ușurință distribuția unui set de date cu multe șabloane. Având acest beneficiu la dispoziție, modelul actual poate fi îmbunătățit dacă este antrenat luând în considerare o cantitate mai mare de triplete de fraze.

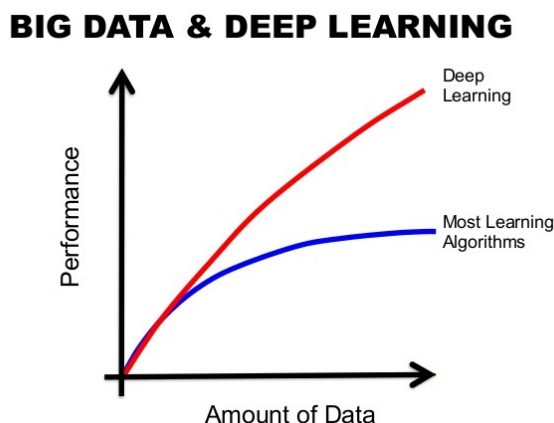


Figura 5.3: Scalabilitatea folosind Machine Learning vs Deep Learning

Desigur, o cantitate mai mare de date nu înseamnă întotdeauna că rezultatul final este mai bun. În urma experimentelor realizate, se poate observa o tendință a chatbot-ului de a fi înclinat către răspunsuri care se regăsesc în setul de antrenare de un număr mare de ori, precum "I'm sorry", "I don't know" etc. Acest fenomen este explicat de faptul că statistic vorbind aceste răspunsuri se apropie foarte mult de media distribuției datelor de intrare. Acest lucru ridică problema diversității cât și a calității datelor care sunt furnizate sistemului. O soluție plauzibilă este antrenarea folosind diverse seturi de date cunoscute ca fiind de referință pentru sarcini de procesare de limbaj natural, precum The Ubuntu Dialogue Corpus, și compararea cu modelul curent.

Capitolul 6

Concluzii

TODO

Bibliografie

- [1] Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. Hierarchical neural network generative models for movie dialogues. *CoRR*, abs/1507.04808, 2015.