EUSKAL HERRIKO UNIBERTSITATEA

Lengoaia eta Sistema Informatikoak

Doktorego-tesia

Hizkuntza-Ulermenari Ekarpenak: N-gramen arteko Atentzio eta Lerrokatzeak Antzekotasun eta Inferentzia Interpretagarrirako

Iñigo Lopez-Gazpio

Donostia, 2018

EUSKAL HERRIKO UNIBERTSITATEA

Lengoaia eta Sistema Informatikoak

Hizkuntza-Ulermenari Ekarpenak: N-gramen arteko Atentzio eta Lerrokatzeak Antzekotasun eta Inferentzia Interpretagarrirako

Iñigo Lopez-Gazpiok Eneko Agirre Bengoaren eta Montse Maritxalar Angladaren zuzendaritzapean egindako tesiaren txostena, Euskal Herriko Unibertsitatean Doktore titulua eskuratzeko aurkeztua.

Donostia, 2018

Nire izebari, Maiteri, inoiz irakurriko ez baduzu ere.

Laburpena

Hizkuntzaren Prozesamenduaren bitartez hezkuntzaren alorreko sistema adimendunak hobetzea posible da, ikasleen eta irakasleen lan-karga nabarmenki arinduz. Tesi honetan esaldi-mailako hizkuntza-ulermena aztertu eta proposamen berrien bitartez sistema adimendunen hizkuntza-ulermena areagotzen dugu, sistemei erabiltzailearen esaldiak modu zehatzagoan interpretatzeko gaitasuna emanez. Esaldiak modu finean interpretatzeko gaitasunak *feedback*a modu automatikoan sortzeko aukera ematen baitu.

Tesi hau garatzeko hizkuntza-ulermenean sakondu dugu antzekotasun semantikoari eta inferentzia logikoari dagokien ezaugarriak eta sistemak aztertuz. Bereziki, esaldi barneko hitzak multzotan egituratuz eta lerrokatuz esaldiak hobeto modelatu daitezkeela erakutsi dugu. Horretarako, hitz solteak lerrokatzen dituen aurrekarien egoerako neurona-sare sistema bat inplementatu eta n-grama arbitrarioak lerrokatzeko moldaketak egin ditugu. Hitzen arteko lerrokatzea aspalditik ezaguna bada ere, tesi honek, lehen aldiz, **n-grama arbitrarioak atentzio-mekanismo baten bitartez lerrokatzeko** proposamenak plazaratzen ditu.

Gainera, esaldien arteko antzekotasunak eta desberdintasunak modu zehatzean identifikatzeko, esaldien interpretagarritasuna areagotzeko eta ikasleei feedback zehatza emateko geruza berri bat sortu dugu: **iSTS**. **Antzekotasun semantikoa eta inferentzia logikoa** biltzen dituen geruza horrekin *chunk*ak lerrokatu ditugu, eta ikasleei feedback zehatza emateko gai izan garela frogatu dugu hezkuntzaren testuinguruko bi ebaluazio-eszenariotan.

Tesi honekin batera hainbat sistema eta datu-multzo argitaratu dira etorkizunean komunitate zientifikoak ikertzen jarrai dezan.

Contributions to language understanding: n-gram attention and alignments for interpretable similarity and inference

Natural language processing can lead to significant improvement of educational applications that are able to reduce the workload of teachers and students. In this dissertation we analyse sentence-level language understanding and make several contributions so that educational systems increase their understanding level, as they are able to process input sentences with greater detail. The fine-grained ability to handle sentences make systems able to produce feedback in learning scenarios.

This dissertation focuses on natural language understanding, and analyses features and systems of both semantic textual similarity and natural language inference. We show that structuring and aligning input in the form of arbitrary word n-grams helps improve results as the modelling capabilities strengthen. We perform our experiments by implementing a state-of-the-art word-level attention based neural network and modify it so that it is able to model and align arbitrary n-grams. Being the alignment between bare words well-known in the recent past, this work presents a large-scale analysis focused on **modelling and aligning arbitrary n-grams**.

Moreover, we add an interpretable layer on top of semantic similarity and language inference to provide educational applications with background to spot the differences and commonalities between a pair of sentences, increase the interpretability of the sentence pair and provide feedback to students. With this new interpretable layer that involves and combines **semantic textual similarity and natural language inference** we are able to align chunks in the sentence pair, explicitly denote commonalities and show that verbalizations in the form of explanations help humans improve accuracy on educational evaluation scenarios.

Several systems and datasets have been released alongside this work so the research community can follow-up research in the field.

Eskerrak

Lehenik eta behin, nire zuzendariei eskerrak eman nahi dizkiet: Eneko eta Montse, eman didaten laguntzagatik. Tesi hau bukatzeko gai ez nintzen izango beraien aholku eta gomendiorik gabe, zalantza uneetan bide egokiak aukeratzen lagundu didatelako eta ikertzen erakutsi didatelako.

Jarraitzeko, IXA taldeari eskerrak eman nahi dizkiot familia txiki honen laguntza uneoro izan baitut tesian zehar izandako arazo guztiei aurre egiteko: teknikoak, kontzeptualak eta emozionalak. Esperientzia zoragarria izan da urte askotan bertan lanean egon izana, are gehiago, alor desberdinetako adituen artean dagoen esperientzia trukea eta laguntzeko prestutasuna guztiz eskertzekoa da.

Azkenik, nire lagunak eta familia eskertu nahi ditut aurrera jarraitzeko behar nuen bultzada izan direlako. Bereziki, nire anaiari, Josuri, zientziaren afizioa ez ezik, zientzia gizartean dibulgatzeko garrantzia erakusteagatik.

Mila-mila esker denoi!

Acknowledgements

Thanks to the Ph.D doctoral grant I have been able to be abroad a total of six months in which I met really incredible people out there. First of all, I'd like to thank Mirella for being my hosting tutor, and for pointing me towards such interesting topics that invaluably contributed towards this dissertation. I also want to thank Siva for his time, interesting comments and help within all the time he guided me while I was in Edinburgh.

Out of the faculty I would like to thank Lynn and Joanna for being my hosts

twice, they always tried to teach me that science can not answer everything and that it is equally meaningful to look inside ourselves for peace, love, enthusiasm and simplicity. I would also like to thank a lot of people I met in the flat that have contributed so much in the personal to me, especially, Flavia from Brazil, Flavia from Italy and Raechel.

Last but not least, I would like to thank and give a huge hug to Donna and Allison because they treat me so nicely since the first day, and because their fish&chips and Sunday roast are the best I've ever had.

A huge hug for all of you, hope we see soon again!

Institutional acknowledgements

This research is supported by a doctoral grant from MINECO (FPU13/00501). It has also been partially funded by MINECO in projects: MUSTER (PCIN-2015-226), TUNER (TIN 2015-65308-C5-1-R), READERS (PCIN-2013-002-C02-01) and SKaTeR (TIN2012-38584-C06-02); by the European Commission in project QTLEAP (FP7-ICT-2013.4.1-610516); and by the Basque Government (A type Research Group, IT344-10). I also thank for technical and human support provided by IZO-SGI SGIker of UPV/EHU and European funding (ERDF and ESF), and, also, I gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU and Pascal Titan X GPU used for this research.

Finally, I would like to thank everyone involved in open software communities for having shared with the world such great tools that make possible works like this one: emacs, latex, org-mode, octave, python, perl, theano, pytorch, tensorflow, weka, scikit-learn, ... I'd like to express my deepest gratitude for their generous effort.

Aurkibidea

1. KAPITULUA

Sarrera

Hizkuntzak ez dira ausaz kateatutako hitz-segida estatikoak, elementu-multzo zabal batez osaturiko sistema konplexuak baizik. Gizakiok sistema konplexu honetaz jabetzeko etengabe gaitasunak garatuz goaz, bi multzotan sailkatu ohi direnak: interpretazioarekin loturiko gaitasunak (hizkuntza-ulermena edo HU) eta sormenarekin loturikoak (hizkuntza-sorkuntza edo HS). Gizakia hizkuntzaz nola jabetzen den azaltzeko hainbat proposamen egin dira hizkuntzalarien artean, eta proposamen sendoenek hiru hizkuntza-oinarri definitzen dituzte: oinarri kulturala, soziala eta biologikoa. Gainera, oinarri horiek aldatzen diren heinean hizkuntzen ibilbidea etengabe aldatuz doala ondorioztatu da, hau da, ingurunetik eta komunikazioen bitartez jasotzen dugun informazioak ez ezik, burmuineko neurona-sareek ere definitzen dute gizakion ahalmen linguistikoa.

Tesi hau hizkuntzaren prozesamenduaren alorrean kokatzen da, makinek hizkuntza ulertzea eta sortzea helburu duen diziplinan. Lan honetan **esaldimailako hizkuntza-ulermena** hobetzeko ataza, sistemak eta datu-multzoak biltzen ditugu. Honetarako, bai hitzak eta hitz n-gramak modelatzeko eta lerrokatzeko, baita hauen arteko interakzioak linguistikoki motibatutako loturen bitartez adierazteko proposamenak egiten ditugu. Interakzioak modu finean modelatzeko eta lerrokatzeko gaitasunak esaldi pare baten inguruko *feedback*a sortzeko aukera ematen digu, era horretan, ikasleen erantzunak kalifikatzea helburu duten sistema adimendunak hobetuz.

1.1 Hizkuntza-ulermena

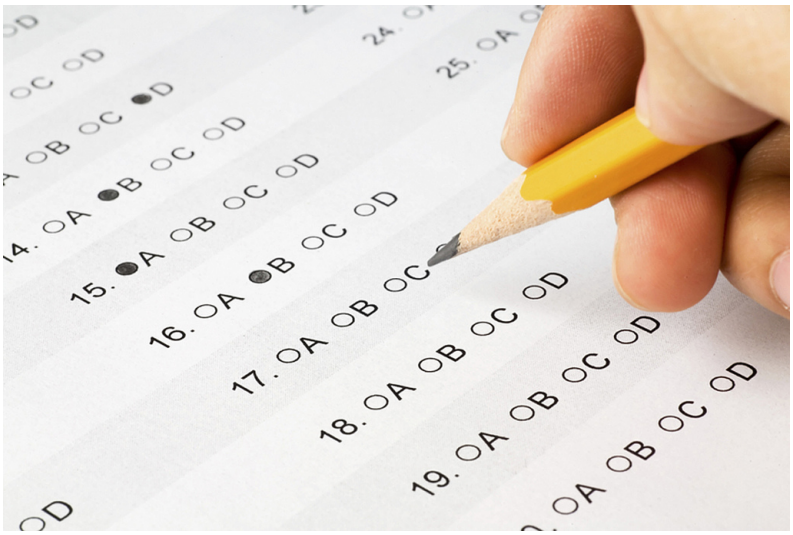
Hizkuntzaren Prozesamendua (HP) informatika, adimen artifiziala eta hizkuntzalaritza biltzen dituen alorra da, hainbat hizkuntza-teknologiaz arduratzen dena (Jurafsky, 2000). HUren helburua **mundua errepresentatzeko eta interpretatzeko** gai diren sistema konplexuak garatzea da, ezagutza eta objektuak modelatzeko errepresentazio abstraktuak makinei helaraziz.

Arazo konplexu honi aurre egiteko asmoz, adituek hainbat teknika eta metodo proposatu dituzte urteetan zehar, besteak beste, ikasketa automatikoan oinarritutakoak, adituen ezagutza erabiltzen duten datu-base, ontologia eta sistema adituak, eta, azken urteetan, ikasketa automatikoaren baitan nagusitzen ari den adar berri bat, neurona-sareetan oinarritutako sistemak, hain zuzen ere.

Hizkuntza ulertzeko gai diren sistemak hainbat aplikazio izan ditzakete, adibidez, irakaskuntzaren alorrean, ikasle batek galdera bati emandako erantzun irekia modelatzea eta erreferentziazko erantzun zuzenaren kontra kalifikatzea. Helburu hau lortzeko, HUren arloak esfortzu handia egin du **teknologian oinarritutako ikasketa-sistemak** garatzeko, *e-learning* aplikazioak, alegia. Adibide gisa 1.2. irudian elektrizitate eta elektronikaren domeinuan ikasleak ebaluatzeko eta ikasketa-prozesua arintzeko diseinaturiko sistema aditu bat ikus daiteke.

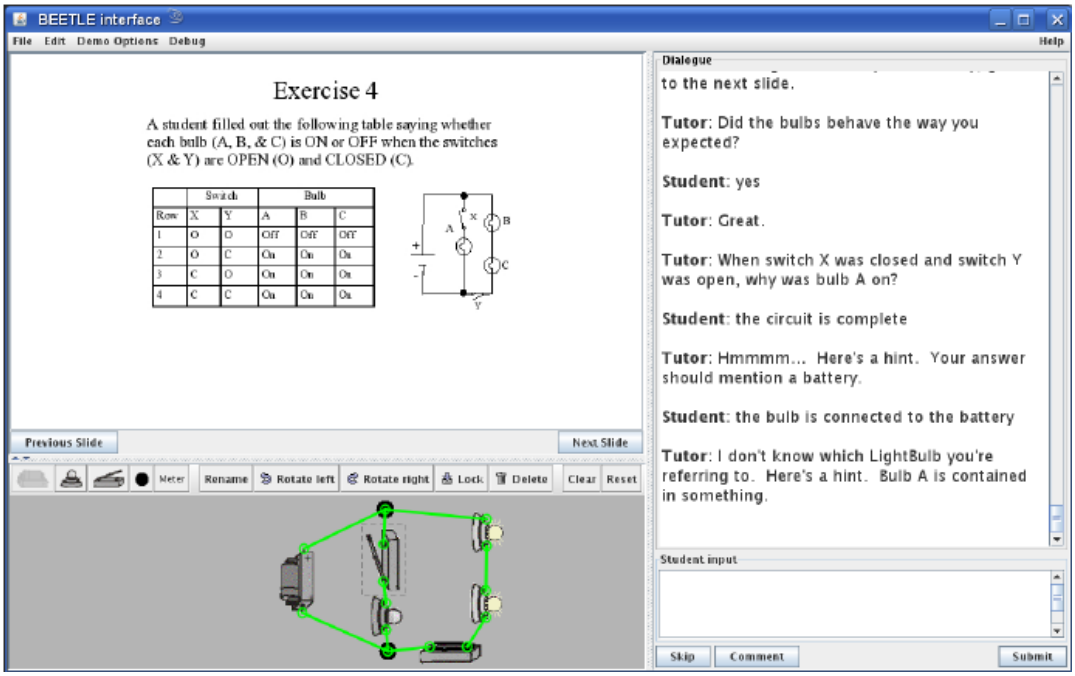
E-learning aplikazioek ikasleen irakurtzeko eta ulertzeko gaitasuna ebaluatzen dute zailtasun-maila desberdinetako testuak eta galderak –irekiak zein itxiak– erabiliz. Galdera irekiak, hau da, hutsetik hasita erantzun beharrekoak formaziorako erabilgarritzat jo dira hainbat lanetan; galdera itxien erabilgarritasuna, ordea, zalantzazkoa da Davies; Conole eta Warburton autoreen arabera (2002; 2005). Aitzitik, galdera itxiak bakarrik jakinduria ebaluatzeko –eta ez asimilatzeko– erabili beharko liratekeela diote hainbat lanek; galdera itxiek ez baitute laguntzen ikasketa-faseetan eduki berriak barneratzen (Karpicke eta Roediger, 2008).

Galdera irekien hezkuntza-balioa handiagoa izanik ere, HUren konplexutasuna dela eta, sistema adituak oraindik ez dira gai ikasleen erantzun irekiak zehatz kalifikatzeko, ez eta ondo edo gaizki erantzundako zatien feedbacka



1.1 irudia: Hainbat lanen arabera galdera itxiak ez dira erabilgarriak eduki berriak barneratzeko. Iturria: `https://www.flickr.com/photos/albertogp123`

emateko ere. 1.2. irudiko sistema adimendunaren dialogoan ikus daitekeen moduan, sistema adituak ikaslearen erantzuna ebaluatzeko arazoak dituenean erreferentziazko erantzun zuzenean agertzen den eta ikaslearen erantzunean agertzen ez den hitz konkretu bat idaztera behartzen du ikaslea. Jokamolde horrekin, konbertsazioa bideratzeko gai da sistema, baina ikasleak idatzitakoaren interpretaziorik egikaritu gabe. Testuinguru honetan, galdera irekiei emandako erantzun askeak erreferentzia baten aurka modu interpretagarri batean ebaluatzeko gai diren sistema adimendunak garatzea oso interesgarria da; hau da, ikasleak idatzitako erantzuna erreferentziazko erantzun zuzenarekin loturak egiteko gai diren sistemak. Lan honetan **interpretagarritasunean** oinarritzen gara, eta honela definitzen dugu termino hau: esaldiko edozein hitz-multzo esplizituki adierazteko eta horri erreferentzia zehatza egiteko gaitasuna izatea. Era horretan, azalpenak ikaslearen erantzunean identifikatutako hitz-multzoekin lotzeko. Sistema adimendunek, interpretagarritasunaz baliatuz, ikaslearen erantzun askea eta erreferentziazko erantzun zuzena aldera dezakete, **esaldiko hitz-multzoak esplizituki lerrokatzeko gai direlako**.



1.2 irudia: BEETLE II tutore-sistema. Iturria: (Dzikovska *et al.*, 2010a).

E-learning aplikazioen esparrua oso alor zabala da, eta, tesi honetan, erreferentziazko esaldi bat eta uneko beste esaldi bat alderatzeko gai diren metodoen inguruan arituko gara. Mota horretako metodoak tutore-sistemetan integratuak egon ohi dira, baita ikaslearen ulermen-maila lantzen duten sistemetan ere. Horrelakoetan, irakaslearen esfortzua murriztea edota akatsen aurrean ikasleari feedback ulergarria itzultzea izaten da motibazio nagusia.

Ikaslearen erantzunak automatikoki ebaluatzeko edo kalifikatzeko sistemek berebiziko aurrerapausoa suposatuko lukete e-learning irakaskuntzan. Esaterako, internet bidez eskaintzen diren ikastaro masiboetan (*Massive Open Online Course* edo MOOC). Testuinguru horietan aditu gutxi batzuk milaka eta milaka ikasleri aurre egin behar diete, eta, guztien galdera irekiak zuzentzea esfortzu handia eskatuko luke. Horregatik, MOOCetan beti galdera itxiak erabiltzen dira, edo ikasleen arteko ebaluazioak. Aldiz, erantzun irekiak interpretagarritasun altuarekin kalifikatzeko gai izango litzatekeen

aplikazio batek laguntza handia emango lioke adituari lan-karga nabarmenki murriztuz, ikasleen akatsen gaineko feedbacka sortuko bailuke.

Hurrengo ataletan ikusiko dugun modura helburu hauek lortzeko **esaldi-mailako antzekotasun eta inferentzia** atazetan sakondu beharko dugu hizkuntza-ulermenaren maila areagotzeko. Antzekotasun semantikoak bi esaldien arteko interakzio semantikoa neurtzen duen balio kuantitatibo bat esleitzea du helburu, aitzitik, inferentzia logikoak kategoria kualitatibo bat esleitzen dio esaldi pareari beren arteko inferentzia-erlazioaren arabera. Aipatutako hurbilpenak esaldien arteko interakzioa islatzea dute helburu, eta gure lanaren abiapuntu izango dira.

1.2 Helburuak eta ikerketa-lerroak

Gaur egun oso ezagunak dira arlo espezifikoak lantzeko sarean atzigarri aurki daitezkeen e-learning aplikazioak. Horiek irakaskuntza, autoikaskuntza edota ikaskuntza ez-presentziala ahalbidetzen dute, eta, gainera, hainbat abantaila eskaintzen dizkiote bai ikasleari baita irakasleari ere; besteak beste, denboraren kudeaketa hobea, irismena zabaltzea, arazo geografikoak ekiditea, zeregin mekanikoak automatizatzea, lan-kargak murriztea eta kooperazio zein elkarlan dinamikak bultzatzea.

Ikasleek galdera irekiei emandako erantzunak automatikoki kalifikatzea eta emandako erantzunaren araberako feedback erabilgarria itzultzea erronka handi bat da HUn. Argi dago ataza horri behar bezala erantzuteko sistema adimendunak ikaslearen erantzuna eta erreferentziazko erantzun zuzenaren arteko antzekotasunak eta desberdintasunak identifikatzeko gai izan behar direla. Ataza bere osotasunean ebaztea batere tribiala ez izan arren, aspalditik da komunitate zientifikoa arazo zehatz batzuei aurre egiteko metodoen bila.

Tesi honen motibazio eta helburu nagusia galdera irekiei dagozkien **erreferentziazko esaldiak eta ikasleek emandako erantzun irekiak kalifikatzeko** sistema hobeak sortzea da, eta, gainera, sistema horiek **feedbacka emateko** gai izatea. Era honetan, hezkuntzaren alorreko HPan pauso bat aurrera egingo genuke.

Ikerketa-lerroak

Tesiko helburu nagusia lortzeko hizkuntza-ulermenean sakondu dugu, egunera arteko atazak eta sistemak aztertuz proposamen berriak egiteko. Zehazki tesi honek bi ikerketa-lerro finkatzen ditu:

1. Esaldi-mailako antzekotasuna eta inferentzia ebazteko gai diren sistemak garatzea.

2. Esaldi-mailako sistema horien interpretagarritasuna areagotzea.

1.3 Ekarpenak

Tesi honen ekarpenak finkatutako bi ikerketa-lerroekin lotzen dira.

Lehen ikerketa-lerroan esaldiak modelatzeko eta errepresentatzeko arkitekturak aztertu ditugu, eta neurona-sareetan oinarritutako sistema berri baten proposamena egin dugu: **hitzen n-grama arbitrarioak modelatzeko eta atentzio-mekanismo baten bitartez lerrokatzeko** gai den sistema. Hitzen arteko lerrokatzea aspalditik ezaguna bada ere, tesi honetan lehen aldiz hitzak baino luzeagoak diren n-grama arbitrarioak kodetzeko eta lerrokatzeko proposamena zabaltzen dugu, eta esaldi-mailako hainbat atazetan ebaluatzen dugu orokortzeko gaitasun ona duela erakutsiz.

Bigarren ikerketa-lerroan **interpretagarritasun altuko HU ataza** berri bat sortu dugu: **iSTS** (*Interpretable Semantic Textual Similarity*). Ataza horretan esaldi parearen arteko antzekotasunak eta desberdintasunak esplizituki adierazita daude. Horretarako, esaldietako *chunk*ak[1] linguistikoki motibatuta lerrokatzen ditugu eta lerrokatzeak antzekotasun eta logika balioekin aberasten ditugu. Irakaskuntzaren alorrari bideratutako ataza honekin batera sortutako **datu-multzoei** esker, sistemek hartutako erabakiak ikasleei azaltzeko gai dira. Konkretuki, irakaskuntzaren alorrean ikasleei feedbacka emateko erabiltzen ditugu azalpenak. Jarraian banan-banan azalduko ditugu tesi honen ekarpenak.

[1]Sintaktikoki motibatutako eta etenik gabeko hitz-segida ez-errekurtsiboak. Definizio hau 4. kapituluan formalizatzen da.

Hitz n-gramen arteko atentzio-ereduak

Hizkuntzalaritza konputazionalaren alorrean hizkuntza-ulermena ahalbidetuko duten sistema adimendunak garatzea da erronka nagusia. Esaldi-mailako atazetan ezagutza errepresentatzeko gai diren sistema hauek esaldi pareak lerrokatu eta esaldien arteko erlazioa adierazten duen irteera-balio bat itzuli behar dute. Antzekotasun semantikoa eta inferentzia logikoa dira erlazioa adierazteko era ezagunenak. Ataza horien zailtasuna eta interesa dela eta, esaldien errepresentazioak sortzeko eta interakzioak modelatzeko gai diren sistemak ebaluatzeko eszenario bilakatu dira. Aipatutako bi atazetan ebaluatzen diren sistema gehienak neurona-sareetan oinarritutako teknologia erabiltzen dute esaldietako kontzeptu abstraktuak bektoreetan kodetzeko, besteak beste, hitz-zakuak, neurona-sare errepikakorrak, neurona-sare errekurtsiboak, konboluzio-sareak eta aurreko teknologiak konbinatzen dituzten sistema-multzoak. Esaldien hitzen arteko interakzioak modelatzeko, ordea, aukera ez da horren zabala eta sistema gehienek hitz bakanetan oinarritzen diren mekanismoak erabiltzen dituzte. Hala eta guztiz ere, sistema onenek atentzio-mekanismoaren erabilera zabala egiten dute. Azken urteetan zalantzarik gabe ondorioztatu baita hitzen arteko interakzioak modelatzeko gai diren atentzio-mekanismoak bereziki interesgarriak direla esaldi pareko loturak modelatzeko.

Neurona-sareak ikasteko eta datuetara doitzeko gaitasun altua duten adimen artifizialeko arkitekturak dira, hierarkikoki antolatu edo kateatu daitezkeenak. Horregatik, geroz eta konplexuagoak diren sistemak garatzeko joera nabarmendu da hizkuntzaren konplexutasuna hobeto ustiatzeko gai direlakoan. Aitzitik, neurona-sareak kateatu eta arkitektura konplexuak garatzeak **hainbat desabantaila** ditu, besteak beste, esaldi pareko hitzen arteko interakzioak neurona-sareen geruzetan zehar sakabanatuta kodetzen direla, interpretagarritasuna galduz. Beste arazo bat da entrenatzeko datumultzo handien beharra dutela –neurona-sareak doitzeko adibide asko behar baitira–. Arazoa ere bada garatzen diren sistemak geroz eta ilunagoak direla –neurona-sareek beren ezkutuko geruzetan sarrerako ezaugarrien errepresentazio geroz eta abstraktuagoak kodetzen dituztelako–.

Esaldi pareen arteko **erlazio semantikoak modelatzeko n-grama arbi-**

trarioak lerrokatzea erabilgarria dela oinarri gisa hartzen dugu lan honetan. Horretarako, hitz bakanetan oinarritutako atentzio-mekanismoak orokortzen ditugu, hitzen ordez hitz n-grama arbitrarioak lerrokatzeko gai diren sistemak garatuz. Gainera, geroz eta ulergaitzagoak diren sistemak garatzeko joeraren aurka, gure proposamena arkitektura ulergarriak eraikitzea da, eta ahal den heinean entrenatzeko datu, denbora eta baliabide gutxi behar dituztenak. Norabide hau jarraitu duten autore gehiago ere badira, eta, lan honetan, autore horiek landutako sistemetan oinarritzen gara proposamen berrien bitartez emaitzak hobetzeko. Hipotesi nagusia da sistema sinpleek ere konplexuen antzera lan egin dezaketela, eta pareko emaitzak lortu daitezkeela arkitektura interpretagarriagoak erabilita.

Atal honetan hitz-zakuetan oinarritutako sistema bat abiapuntu gisa hartuta **n-grama arbitrarioak lerrokatzeko gaitasuna** ematen diogu. Hitz-zakuetan (*Bag-of-Words* edo BoW) oinarritutako sistemen arazo nagusia hitzak elementu independente gisa tratatzen dituztela da. Beraz, testuingurua gehitzean, mota horretako sistemen eraginkortasuna areagotzea posible da. Kontrara, proposatzen dugun sistemak sarrerako esaldi pareen n-grama arbitrarioen errepresentazioak modelatzen, eta, errepresentazio horien gaineko lerrokatzeak bere kabuz egiten ikasten du. Horretarako, lehenbizi esaldi bakoitzeko ondoz ondoko hitzen sekuentziak erauzten ditu eta n-gramak esplizituki lerrokatuta mantenduko dituen matrize bat osatzen du, n-grama arbitrarioen gaineko atentzio-mekanismoa, alegia.

N-gramak lerrokatzeko proposatzen dugun arkitektura HUko bi ataza desberdinetan ebaluatzen dugu: antzekotasun semantikoan eta inferentzia logikoan, guztira bost datu-multzo desberdinetan. Jatorrizko BoW arkitekturarekiko eta alderatzeko proposatzen ditugun hedapenekiko (*baseline*) lortutako hobekuntzek gure proposamenaren lerro nagusia frogatzen dute: egitura n-grama bidezko lerrokatzeen bitartez hornituz BoW motako sistemak hobetzeko gaitasuna dagoela. Sare errepikakorren, errekurtsiboen eta konboluzio-sareen bitartez hedatutako baselinen kontra ebaluatzen dugu gure proposamena, eta eszenario guztietan gure sistemak emaitza hobeak lortzen dituela azaltzen dugu. Horrek, n-gramak lerrokatzea beste alternatibak baino aukera hobea izan daitekeela erakusten du, etorkizunerako bideak zabalduz. Pro-

posatutako arkitektura aurrekarietan atzigarri dauden sistemekin egindako alderaketan ere pareko emaitzak lortzen ditugula argudiatzen dugu, gure sistemaren eraginkortasuna bermatuz. Gainera, bereziki HUren maila ebaluatzeko diseinaturiko datu-multzoen adar zailetan ere emaitza onak lortzeko gai gara, gure sistema **datu-multzo txikietarako bereziki aproposa** dela erakutsiz.

Bukatzeko, aipatu oinarri gisa hartutako BoW sistemak hitzen –eta gure hedapenaren bitartez n-gramen– arteko interakzioak esplizituki kodetzeko gaitasuna duenez, sistemak hartutako erabakiak interpretatzeko bidea zabaltzen digu etorkizunari begira. Orain n-gramen interakzioak eskalar baten bitartez adierazita egonik ere, etorkizunean lerrokatzeak aberastea litzateke helburua, lerrokatze bakoitza anotazio linguistikoez hornituz.

Horrekin guztiarekin, gure tesiko lehen ikerketa-lerroa bermatzen dugu: hitzak baino luzeagoak diren n-grama arbitrarioak lerrokatuz erlazio semantikoak eraginkortasunez modelatzeko gai diren sistemak inplementatu baititugu.

Esaldien arteko desberdintasunak topatzen eta azaltzen

Esaldi-mailako HU atazek oinarrizko eszenario bat definitzen duten heinean, irakaskuntzan sistema adimendunen ulermen-maila eta inferentzia-gaitasuna haratago eramatea eskatzen da. Ikasleei feedback zehatza emateko ikasleak ebaluatu ez ezik, akatsei dagokien azalpen esanguratsuak ere eman behar direlako. Ulermen-maila areagotzeko esaldi-mailakoak baino zehatzagoak diren kontzeptuak modelatu behar dira, eta esaldi barneko kontzeptuak **interpretagarri bilakatu**.

Gure lanaren hipotesi nagusia sistema adimendunek **bere burua azaltzeko** beharra dutela da, gizakioi azalpen ulergarriak eman nahi badizkigute behintzat. Helburu horrekin, eta irakaskuntzaren alorrean esaldi-mailako HU areagotzeko motibazioarekin antzekotasun semantikoa eta inferentzia logikoa ataza bakar batean uztartzen ditugu. Horretarako, **interpretagarria** den geruza berri bat definituz: iSTS edo antzekotasun semantiko interpretagarria. iSTS erabiliz edozein esaldi pareko hitzak lerrokatzea posible da,

eta lotura bakoitzari antzekotasun semantikoaren balio bat eta inferentzia logikoaren kategoria bat esleitzea (ikus 4.1 Irudia). Geruza berri horrekin ikasleei feedback zehatza emateko gai izan gara, eta, ondorioz, hezkuntzaren alorreko hizkuntzalaritza konputazionalean pauso bat aurrera egin dugu. iSTS bi urtez egon da aktibo SemEvaleko workshopean, 2015 eta 2016 urteetan hurrenez hurren (Agirre *et al.*, 2015a, 2016). Urte horietan zehar hainbat datu-multzo eta anotazio-gidalerro eskuragarri jarri ditugu komunitate zientifikoan, eta hainbat sistema garatu eta ebaluatu dira.

2015. urtean ataza **prototipo**[2] gisa plazaratu genuen *STS* izeneko antzekotasun semantikoaren atazarekin batera. Bertan ataza bera, anotazio-gidalerroen lehen bertsioa, sistemak entrenatzeko zein ebaluatzeko datu-multzoak eta ebaluazio-eszenarioak definitu ziren. Sistemen helburua esaldi pareak emanik pareen chunkak identifikatzea zen, eta horiek identifikatutakoan chunken arteko loturak egitea. Lotura bakoitzari antzekotasun-balio bat eta inferentzia logikoari dagokion kategoria bat esleitu behar zieten sistemek, eta, hain zuzen ere, burututako elkar-lotzeen zein esleitutako balio zein kategorien zuzentasunaren arabera ebaluatu ziren sistemak. Datu-multzoei dagokienez bi domeinutako ikasketa- eta ebaluazio-multzoak atzigarri jarri ziren: irudien laburpenetan oinarritutakoa bata, eta egunkarietako berrien izenburuena bestea.

Deskribaturiko atazan parte hartzeko asmoz ur-jauzi diseinua jarraitzen zuen arkitektura bat garatu genuen (Agirre *et al.*, 2015b). Arkitektura lau osagaiz baliatzen da iSTS burutzeko: (1) esaldietan chunkak identifikatzeko azaleko *parserra*, (2) chunkak elkar lotzeko azaleko sintaxian oinarritutako lerrokatzailea, (3) antzekotasun semantikoari dagozkion balioak esleitzeko ikasketa automatikoan oinarritutako erregresoreak eta (4) inferentzia logikoari dagozkion kategoriak esleitzeko ikasketa automatikoan oinarritutako sailkatzaileak.

Sistema horrek aurrekarien egoerako emaitzak lortu zituen atazan. Gainera, atazan ez ezik, **irakaskuntzaren domeinuko ebaluazio bitan** erabili da sistema: iSTSren erabilgarritasuna bermatzeko diseinaturiko bi esperimentutan hain zuzen ere. Lerrokatzeak berbalizatzeko gai den algoritmo sinple baten bitartez iSTS estatistikoki esanguratsua dela frogatu dugu ira-

[2] `http://alt.qcri.org/semeval2015/task2/index.php`

kaskuntzaren domeinuko esperimentuetan, bietan iSTSren bitartez ikasleei feedbacka ematea lagungarria dela erakutsiz.

iSTSren erabilgarritasuna eta izandako parte-hartze altua bermatuta 2016. urtean ataza bere gisara plazaratu genuen SemEvalen[3] bigarren aldiz. Urte horretan aurreko urteko ataza pilotoa findu, datu-multzoak hedatu, eta, gainera, irakaskuntzaren domeinuko datu-multzo berri bat plazaratu genuen. 2016. urteko iSTS atazan parte hartzeko aurreko urteko sistema hartu eta ikasketa automatikoan oinarritutako ereduak neurona-sareetan oinarritutako eredu berriengatik ordezkatu genituen. Neurona-sare errepikakorretan oinarritutako sistema hau sarrera gisa jasotako chunken errepresentazio abstraktuak sortzeko gai zen, eta, horretaz gain, antzekotasun semantikoari zein inferentzia logikoari dagozkien balioak eta kategoriak batera esleitzeko. Horretarako, errepresentazio abstraktuak sortzeko gai diren neurona-sare errepikakorrez gain, beste bi neurona-sare ere erabiltzen zituen: bata erregresioa egiteko, eta, bestea, sailkapena egiteko. Neurona-sareekin aurreko urteko emaitzak hobetu ez ezik, irakaskuntzako datu-multzoan atazako emaitzarik onenak lortu genituen (Lopez-Gazpio *et al.*, 2016b).

Ekarpen hauekin tesiko bigarren ikerketa-lerroa bermatzen dugu: esaldimailako sistemen interpretagarritasuna areagotu baitugu, eta bere erabilpena hezkuntzaren arloko testuinguruetan ebaluatu.

[3] `http://alt.qcri.org/semeval2016/task2/`

1.4 Tesiaren egitura eta osatzen duten argitalpenak

Tesia argitalpenen bilduma gisa aurkeztu da, eta euskaraz zein ingelesez idatzita dago, jarraian adierazten den eran antolatuta: hiru kapitulu euskaraz eta bi kapitulu ingelesez, azken hauetako bakoitza argitalpen banaz osatua.

Ingelesez idatzitako bi kapituluek tesiaren muina osatzen dute, atal teknologikoena, alegia, eta formatu aldetik maketatu egin dira tesiaren egitura orokorra mantentzeko. Jarraian tesiko kapitulu guztiak azaltzen ditugu:

1. Lehen kapituluan, sarreran, tesia kokatu eta lan esparrua zehazten dugu, besteak beste, tesian zehar bereziki garrantzitsua den interpretagarritasunaren kontzeptua plazaratzen dugu.

2. Bigarren kapituluan aurrekariak deskribatzen ditugu, bertan, HUko aplikazioak, neurona-sare arkitekturak eta esaldi-mailako antzekotasun semantikoa zein inferentzia logikoa sakonean deskribatzen dugu.

3. Hirugarren kapituluan tesiaren muina diren oinarrizko hizkuntza-ulermen atazak deskribatzen dira zehatz-mehatz, eta, baita, oinarrizko ataza horietan egindako proposamen berriak ere. *Knowledge-Based Systems* aldizkarira bidalitako artikulu batek osatzen du kapitulua.

   Word n-gram attention models for sentence similarity and inference

Iñigo Lopez-Gazpio, Montse Maritxalar, Mirella Lapata and Eneko Agirre.
Word n-gram attention models for sentence similarity and inference
Preprint submitted to Knowledge-Based Systems.
ISSN: 0950-7051

4. Laugarren atalean antzekotasun semantiko interpretagarria plazaratzen dugu, *Knowledge-Based Systems* aldizkarian argitaratutako lan batek osatzen du kapitulua.

   Interpretable semantic textual similarity: Finding and explaining differences between sentences

Iñigo Lopez-Gazpio, Montse Maritxalar, Aitor Gonzalez-Agirre, German Rigau, Larraitz Uria and Eneko Agirre.
Interpretable semantic textual similarity: Finding and explaining differences between sentences.
Knowledge-Based Systems. 119, pp. 186 - 199. Elsevier.
ISSN: 0950-7051, Impact Factor: 3.325.
DOI: http://dx.doi.org/10.1016/j.knosys.2016.12.013, 2017.

5. Azken kapituluak, bosgarrenak, tesiaren ondorio eta etorkizuneko lanak biltzen ditu.

Lotutako beste argitalpenak

Atal honetan tesiarekin erlazionatutako gainerako argitalpenak zerrendatzen ditugu, tesiarekin duten loturaren azalpenarekin batera. Hainbat lanetan autoreak alfabetikoki ordenatuta agertzen dira.

1. **SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability**

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel M Cer, Mona T Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar and Rada Mihalcea.
SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability.
Proceedings of the 9th International Workshop on Semantic Evaluation.
pp. 252 - 263. SemEval NAACL-HLT.
ISBN: 978-1-941643-40-2, 2015.

Artikulu honetan antzekotasun semantikoari dagokion atazaren antolakuntza deskribatzen da, eta baita, azpiataza gisa, antzekotasun semantiko interpretagarriari dagokion ataza pilotoa ere. Antzekotasun semantiko interpretagarria lehen aldiz publikatuta agertzen da artikulu honetan, parte-hartze eta interes handia erakutsiz komunitate zientifikoan. Ataza pilotoarekin batera parte-hartzaileentzat prestatutako datu-multzoak, baselineak, parte-hartzaileen sistemak eta lortutako emaitzak plazaratzen dira.

2. **UBC: Cubes for English Semantic Textual Similarity and Supervised Approaches for Interpretable STS**

Eneko Agirre, Aitor Gonzalez-Agirre, Iñigo Lopez-Gazpio, Montse Maritxalar, German Rigau and Larraitz Uria.
UBC: Cubes for English Semantic Textual Similarity and Supervised Approaches for Interpretable STS.
Proceedings of the 9th International Workshop on Semantic Evaluation.
pp. 178 - 183. SemEval NAACL-HLT.
ISBN: 978-1-941643-40-2, 2015.

Antzekotasun semantikoari, eta, baita, antzekotasun semantiko interpretagarriari dagozkien atazetan parte hartzeko eraikitako sistemak deskribatzen dira lan honetan, baita sistemek lortutako emaitzak ere. Bertan, antzekotasun semantiko interpretagarria ebazteko hartutako urratsak deskribatzen dira, ezaugarrietan oinarritutako ikasketa automatikoko ereduak erabilita. Sistema honek aurrekarien egoerako emaitzak lortu zituen atazako datu-multzoetan.

3. **SemEval-2016 Task 2: Interpretable Semantic Textual Similarity**

Eneko Agirre, Aitor Gonzalez-Agirre, Iñigo Lopez-Gazpio, Montse Maritxalar, German Rigau and Larraitz Uria.
SemEval-2016 Task 2: Interpretable Semantic Textual Similarity.
Proceedings of the 10th International Workshop on Semantic Evaluation.
pp. 178 - 183. SemEval NAACL-HLT.
ISBN: 978-1-941643-95-2, 2016.

Artikulu honetan antzekotasun semantiko interpretagarriaren ataza deskribatzen da bere bigarren plazaratzean. Bigarren aldi honetan, aurreko aldiarekiko hainbat hobekuntza deskribatzen dira. Berrikuntza garrantzitsuenak ataza pilotoaren chunkak lerrokatzeko mugak gainditzea eta irakaskuntzaren alorreko datu-multzo berri bat gehitzea dira. Datu-multzo berri bat gehitu ez ezik, aurretik plazaratutako datu-multzoak ere zabaltzen dira.

4. **iUBC at SemEval-2016 Task 2: RNNs and LSTMs for interpretable STS**

Iñigo Lopez-Gazpio, Eneko Agirre and Montse Maritxalar.
iUBC at SemEval-2016 Task 2: RNNs and LSTMs for interpretable STS
Proceedings of the 9th International Workshop on Semantic Evaluation.
pp. 771 – 776. SemEval NAACL-HLT.
ISBN: 978-1-941643-95-2, 2016.

Neurona-sareetan oinarritutako sistemak deskribatzen dira artikulu honetan, antzekotasun semantiko interpretagarria burutzeko eraikitakoak. Sistema horiek aurretik ezaugarri-zerrendetan oinarritutako ereduak gainditzen dituzte, eta, gainera, atazak berrikuntza gisa duen irakaskuntzaren alorreko datu-multzoan emaitza onenak lortzeko gai dira, aurrekarien egoerako muga berri bat zehaztuz.

5. **Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual**

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio and Lucia Specia.
Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation
Proceedings of the 11th International Workshop on Semantic Evaluation.
pp. 1 – 14. SemEval NAACL-HLT.
ISBN: 978-1-945626-55-5, 2017.

Artikulu honetan antzekotasun semantikoari dagokion atazaren antolakuntza deskribatzen da. Bertan ataza bera, parte-hartzaileentzat prestatutako datu-multzoak, baselineak, parte-hartzaileen sistemak eta lortutako emaitzak plazaratzen dira.

6. **Erantzunen kalifikazio automatikorako lehen urratsak**

Eneko Agirre, Itziar Aldabe, Oier Lopez de Lacalle, Iñigo Lopez-Gazpio and Montse Maritxalar.
Erantzunen kalifikazio automatikorako lehen urratsak.
EKAIA Euskal Herriko Unibertsitateko Zientzi eta Teknologi Aldizkaria. 29.
ISSN: 0214-9001.
DOI: 10.1387/ekaia.14530, 2015.

Antzekotasun semantikoa eta inferentzia logikoa aztertzea du helburu artikulu honek. Bertan bi teknika hauen analisia eta irakaskuntzaren alorrean aritzeko dituzten dohainak eta mugak deskribatzen dira.

1.5 Aurrekariak Ixa taldean

Aurkezten den tesi hau ez da ezerezetik eratu den lan bat. Tesi honen ikerketa Ixa[4] taldean egin da, eta talde honetako beste tesi bat aipatzea beharrezkoa da lan hau kokatzeko. Ixa taldeak Euskal Herriko Unibertsitatean (EHU) hizkuntzaren prozesamenduan dihardu lanean, eta bertan antzekotasun semantikoaren ataza sortu zen (Agirre *et al.*, 2012). Denboran zehar zabaldu ez ezik, kontsolidatu ere egin da eta parte-hartze zein onespen handia izan du. Aipatutako atazaren inguruan aurrekaria den (Gonzalez-Agirre, 2017) tesia dugu aurrekari. Gainera, aipatutako lan horien autoreen laguntza handia izan genuen tesi honen hasierako pausoak zehaztu genituenean, guztiak Ixa taldeko ikerlariak. Tesi honetan antzekotasun semantikoa haratago eramaten dugu hitzen n-gramen lerrokatzeak egikarituz eta ataza bera inferentzia logikoarekin uztartuz.

[4] `http://ixa.si.ehu.es/`

2. KAPITULUA

Aurrekariak

Adimen artifizialaren helburua konputagailuei gizakion pareko inferentzia-gaitasuna ematea da. Zientziaren esparru hori lehen arrastoetatik gaur egunera arte asko aldatu da, konputagailuak "hiru lerrokatu" moduko jokoak jolasteko trebatzetik gidaririk gabeko autoak gidatzera, edo milioika pertsonek erabiltzen dituzten sare sozialak metodo automatikoen bitartez kudeatzera. Azken alor honetan eragin handia dute hizkuntzaren azterketa eta prozesamenduko teknologiek komunikazio gehientsuenak testu bidez burutzen baitira, hala nola: whatsapp edo telegram mezuak, korreo elektronikoak, tweetak eta egunkari elektronikoak.

Kapitulu honetan HUren aurrekarietan sakonduko dugu, lotutako atazak eta sistemak deskribatuz. Bereziki, esaldi pareak modelatzea helburu duten atazak azalduko ditugu, beren indarguneak, ahuleziak eta mugak identifikatuz, era horretan, tesi honen abiapuntua zehaztuz. Ikusiko dugun moduan, esaldimailako atazak hizkuntza-ulermenean gertatzen diren arazo konplexuagoei irtenbidea emateko giltza izango dira.

2.1 Hizkuntza-Ulermena

HUk testua irakurri eta ulertzeko gai diren sistema adimendunak trebatzea du helburu. Alor honi, ingelesez, "machine reading" deritzo eta hainbat ataza biltzen ditu, esaterako: dokumentu bilduma batetik informazio zehatza erauztea, dokumentu bilduma batetik galdera jakin bat modu optimoenean erantzuten duten dokumentu egokienak erauztea edo dokumentu bat oinarri gisa hartuta galdera jakin batzuen erantzunak topatzea. Aplikazio mota honen adibide garbia eta ezaguna dugu IBM Watson (High, 2012), Jeopardy! jolasteko trebatutako adimen artifiziala.

Garbi dago ataza hauek guztiak behar bezala burutzeko hizkuntza-ulermen eta kontzeptuen errepresentazio-maila altua behar dela; aitzitik, ataza gehienak dokumentu bat edo dokumentu batetik informazioa bere horretan erauztea eskatzen duten heinean, irakaskuntzako hizkuntza-ulermenerako atazek sistema adimendunen ulermen-maila eta inferentzia-gaitasuna haratago eramatea eskatzen dute: ikasleak ebaluatzeko eta akatsei dagozkien azalpen esanguratsuak emateko. Horretarako, kontzeptuak modelatzeko errepresentazio abstraktuak kudeatu ez ezik, interpretagarriak ere izatea beharrezkoa da. Irakaskuntzaren alorrean hainbat aplikazio aurkitu ditzakegu, besteak beste; idazleei laguntza ematen dieten aplikazioak (Macdonald *et al.*, 1982); idazlanen inguruko gramatika, diskurtso mekanismoak eta idazketa-estiloak ebaluatzen dituzten aplikazioak (Burstein *et al.*, 2013); hizkuntzak ikasten ari diren ikasleei errore kontzeptualak idazlanetan identifikatzen laguntzen dieten aplikazioak (Leacock *et al.*, 2014); edo dokumentu baten gainean proposatutako galdera irekiei emandako erantzunen osotasuna eta zuzentasuna ebaluatzeko aplikazioak (Burrows *et al.*, 2015).

Hezkuntzaren alorreko aplikazioak garatzeko hainbat hurbilpen erabili dira denboran zehar, hainbat muga agerian utzi dituztenak. Hasiera batean, domeinu espezifikoetan lan egiteko sistemak garatzeko joera nabarmendu zen (Callaway *et al.*, 2006; Dzikovska *et al.*, 2010a). Sistema haietan domeinu jakin bateko ezagutza erauztea zen ideia nagusia, eta, era horretan, eskuz kodetutako erregelak erabilita domeinu horretan inferentziak eta arrazonamendu konplexuak egiteko gai ziren. Hurbilpen horiek domeinu jakin baten barruan arrazoitzeko gaitasun handia lor zezaketela erakutsi zuten, aitzitik,

erregeletan oinarritutako sistemak mantentzea eta eskalatzea oso lan astuna da, eta, horregatik, domeinuak oso esparru txikia estal zezakeela ondorioztatu zen, zehaztasun handiko arrazoibidea lortu nahi bada behintzat. Gaur egun, estaldura zabaltzeko asmoz, ikasketa automatikoan oinarritutako sistemak erabiltzeko hurbilpenak nabarmendu dira. Ikasketa automatikoa erabiltzeko motibazio nagusia da domeinutik domeinura aldatzeko esfortzu txikiagoa egin behar dela, erregeletan oinarritutako sistemekin alderatuz gero –datu berriak biltzea besterik ez–.

Teknika bakoitzak hainbat abantaila eta desabantaila ditu, esaterako: adituen ezagutza erabiltzen duten teknikek zehaztasun handia izan ohi dute, baina, aldi berean, domeinuaren estaldura txikia. Izan ere, ezagutza edo kontzeptuak datu-baseetan zein ontologietan txertatzea oso eragiketa garestia da, bai ezagutza berria identifikatzeko beharragatik, baita osotasunarekiko trinkotasuna mantentzeko eskatzen duen esfortzuagatik ere (Álvez *et al.*, 2018). Beste alde batetik, estatistika zein ikasketa automatikoan oinarritzen diren teknika eta metodoek datuekiko dependentzia handia dute, hau da, informazio-iturri bat behar dute eredu edo patroiak ikasteko. Aldiz, behin eredu edo patroi hauek ikasita kasu berrietara orokortzeko gaitasun handia dute, eta, hori, oso ideia interesgarria da ezagutza-iturri finitu bat nahi adina estrapolatu edo orokortu baitaiteke. Teknika horien arazo nabari bat da datu-multzo erraldoiak behar dituztela, eta halako baliabideak sortzea zeregin garestiak izan ohi dira. Gainera, datu-multzoak handitzen diren heinean hauen kalitatea bermatzea geroz eta zailago bihurtzen da.

2.2 Neurona-sareak eta espazio semantikoak

Azken urteetan neurona-sareetan oinarritutako metodo konputazionalek nabarmenki aurreratu dute alor desberdinetako egoera. Hizkuntzaren azterketa eta prozesamenduan ez ezik, hizketaren prozesamenduan, robotikan, konputagailu bidezko ikusmenean, biomedikuntzan eta beste arlo askotan ere emaitzak modu esanguratsuan hobetzea lortu da (Young *et al.*, 2017). Metodo hauen arrakasta, ospea eta ahalmen teknologikoa azaltzeko lau faktore nagusi aipatu daitezke.

Lehen faktorea, datuetan oinarrituta egotea. Orain arte ezagututako ikasketa automatikoko sistema konbentzionalek gizakion aldetik esfortzu eta domeinuaren ezagutza handia eskatzen zuten ezaugarri esanguratsuak sortzeko. Ezaugarri esanguratsuetatik abiatuta sistema horiek gai ziren sailkapen edo erregresio patroiak ikasteko. Neurona-sareen indarguneetako bat da –ikasketa automatiko tradizionalaren aldean– datuen errepresentazio abstraktuak automatikoki ikasteko gaitasuna dutela, geruzetan antolatu daitezkeen neurona-sareen kateaketak euskarri gisa erabilita.

Bigarren faktorea, datu-iturri erraldoiak atzigarri izatea. Neurona-sareak entrenatzeko eta pisuak optimizatzeko datu kopuru handiak behar dira. Datuak erregistratzeko egungo egoera teknologikoa optimoa da metodo horiek entrenatzeko beharrezko diren datuak biltegiratzeko.

Hirugarren faktorea, hardwarearen aurrerapen teknologikoa. Neurona-sareetan erabiltzen diren teknikak aspalditik ezagunak dira (Rumelhart *et al.*, 1986). Garai hartako teknologia konputazionala sistema konplexu horiek entrenatzeko gai ez bazen ere, egun, GPUen (*Graphical Processing Unit*) konputazio-ahalmenarekin arazo hau gainditu da, eta eredu konplexuak entrenatzea posible da. GPUei esker konputazio-ahalmen masiboa denbora finitu batean aurrera eramateko gaitasuna dago.

Laugarren faktorea, baliabide teknikoen atzipena. Software libreko edota itxiko baliabideei esker neurona-sareetan oinarritutako ereduak ez ezik esperimentuak azkar diseinatzeko lanak asko erraztu dira. Burututako esperimentuak eta lorpen berriak berriro publikatu eta komunitate zientifikoan atzigarri uzteko joera hau uneko garapen azkarraren erantzule da.

Faktore horiek guztiak direla eta **neurona-sareak** ikerkuntzan uneoro erabiltzen ari dira ikertzaileak, eta, gainera, erabilera geroz eta gehiago zabaltzen ari da alor desberdin askotan, ondo orokortzeko gaitasuna erakusten ari direlako. Gaur egun, neurona-sareez osatutako sistemak gizakion mailako zehaztasuna eta errendimendua izatera iritsi dira hainbat arlotan, bereziki aipatu daitezke hedapen mediatiko nabarmena izan duten auto autonomoak eta munduko GO jokalari onenak irabazteko gai izan diren adimen artifizialak (Churchland eta Sejnowski, 2016).

Arrakasta honen ondorioz hizkuntzaren prozesamenduan aspalditik ezagunak diren azaleko sintaxitik zein testutik erauzitako ezaugarriak erabiltzen dituzten ikasketa-metodoen erabilera gutxitu egin da. Uneko joera neurona-sareen kateaketetan oinarritzen diren sistema konplexuak erabiltzea da, eta, sarrerako testuen errepresentazio abstraktuak ikasteko geroz eta gehiago nagusitzen ari diren **hitz-bektoreak**.

Hitz-bektoreak hitzak, hitzen egiturak eta hitzen arteko erlazioak matematikoki deskribatzeko edo modelatzeko erabiltzen diren eskalarren bektoreak dira (Mikolov *et al.*, 2013b). Ezagutza modelatzeko teknika horrek emaitza onak erakutsi baditu ere, ez da HPn ezagutza errepresentatzeko teknika bakarra. HUren alorrean semantikaren azpiatala da hitzen errepresentazioak aztertzen eta egituratzen saiatzen dena, eta, denboran zehar proposamen desberdinak aurkeztu ditu ulermen abstraktu hau errepresentatzeko eta ebaluatzeko. Egia esan, mundua errepresentatzeko teknika onenak identifikatzeko ez ezik, ezagutza semantikoa nola antolatu beharko litzatekeen erabakitzeko ere eztabaida handia dago.

Hasiera batean esaldiak semantikoki modu esanguratsuan deskribatzeko metodo sinbolikoak erabili baziren ere, gaur egun ezagutza modelatzeko sare semantikoak eta **metodo distribuzionaletan** oinarritutako espazio semantikoak dira gailendu direnak. Sare semantikoek (Collins eta Quillian, 1969) grafoen bitartez islatzen dute ezagutza, nodoek kontzeptuak osatzen dituzte eta nodoen arteko konexio edo ertzek erlazio semantikoak. Mota honetako ezagutza-baseen artean aurkitu daitezke: WordNet, FrameNet, VerbNet, PropBank, Yago, etab. Lan honetan **espazio semantikoak** (Wittgenstein, 1953) lantzen ditugu. Espazio semantikoetan hitzen esanahiek es-

zenario linguistikoak zehazten dituztela aurresuposatzen da, eta, hipotesi distribuzionalari jarraiki (Harris, 1954) esanahi antzeko hitzak testuinguru berean agertzeko joera dutenez, hitz-bektoreak testuingurutik ikasten dira metodo ez-gainbegiratuak erabilita. Metodo honen arabera Hilberten espazio n-dimensional batean antzeko testuingurua duten hitzak gertu egongo dira elkarren artean, eta zerikusirik ez duten hitzak, ordea, urruti (ikus 2.1 Irudia). Espazio semantikoak neurona-sare sistemekin konbinatzen direnean bikote paregabea bihurtzen dira, izan ere, neurona-sareetan oinarritutako sistemak espazioko bektoreetan kodetutako ezagutzatik ezaugarri abstraktuak eraucteko gai dira. Collobert *et al.* (2011) autoreek neurona-sareetan eta espazio semantikoetan oinarritutako arkitekturekin, inolako eskuzko ezaugarri gehigarririk erabili gabe, alor askotan ordura arte ezagututako emaitzak hobetu zituzten konbinazio honen –neurona-sareen eta hitz-bektoreen– egokitasuna azpimarratuz.

Semantikako adar desberdinetako ezagutza batera errepresentatzeko ahaleginak ere egin dira aurrekarietan, esate baterako, Lewis eta Steedman (2013) autoreek semantika distribuzionala eta lehen mailako logika konbinatzeko saiakera egin zuten. Helburua semantika distribuzionalaren onurak –hitzen errepresentazio abstraktu esanguratsuak– eta lehen mailako logikaren onurak –inferentzia-gaitasuna– konbinatzea zen, HUri begira errepresentazio semantiko osoagoak eraikitzeko. Horretarako, espazio semantikoetan oinarrituta, errepresentazio logikoak batera mapatzen saiatzen dira metodo ez-gainbegiratuen laguntzarekin. Beste alde batetik, semantika eta logika konbinatzen saiatu den beste adar bat aipatu daiteke, logika probabilistikoa (Beltagy *et al.*, 2014). Lan horien motibazio nagusia logikako inferentzia-erregelen zurruntasuna samurtzea da, horretarako, inferentzia-erregelei probabilitate pisuak esleituz.

Ezagutza-baseak eta lehen mailako logika-inferentziak gai interesgarriak izan arren, lan honetan semantika distribuzionaleko espazio semantikoak aztertuko ditugu sakon. Ikerketa honen hipotesi nagusiari jarraituz, **esaldi-mailako HU landuko dugu** jatorrizko esaldiak oinarritzat hartuz, eta, horretarako, hitzen espazio semantikoak, hitz-bektoreak eta hauen gaineko konposizionaltasuna lantzea beharrezkoa da. Ezagutza-baseak hitz-bektoreak hedatzeko

2.1 irudia: Hitz eta kontzeptu-multzo desberdinei dagozkien hitz-bektoreak espazioan proiektatuta. Iturria: "Multimodal Deep Learning winter school 2017". Universitat Politècnica de Catalunya. `https://telecombcn-dl.github.io/2017-dlsl/`

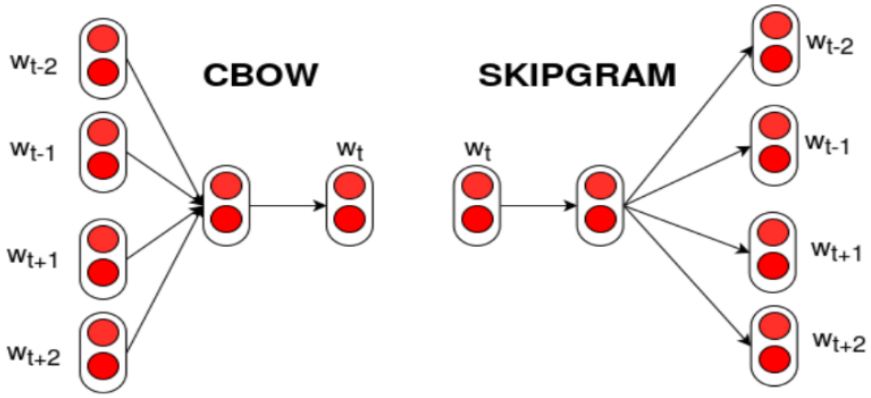
erabili badira ere (Goikoetxea *et al.*, 2016), etorkizuneko lanetarako uzten dugu azterketa hori.

Hitz-bektoreen alde aipatu beharra dago neurozientzietan egindako hainbat lanen arabera espazio semantiko distribuzionalen eta gizakion burmuineko errepresentazioen artean antzekotasunak aurkitu direla. Alde batetik, kontzeptuen errepresentazioak burmuineko neurona-sareen aktibazioekin lotuta daudela azpimarratu da, eta, ezagutza hori bektore gisa kodetzeko aukera dagoela (Haxby *et al.*, 2001). Beste alde batetik, hitz-bektoreak ikasteko erabiltzen diren corpusetan oinarritutako metodoak burmuineko aktibazio-patroiak aurresateko gai direla erakutsi da, pertsona bat kontzeptu batean pentsatzen dagoen bitartean behintzat (Murphy *et al.*, 2012).

Hitz-bektoreak hitzak espazioan ezaugarritu ditzaketen eskalarren bektoreak dira. Bektore bakoitzak hitz bat irudikatzen du espazioan, eta, horiek ikasteko erabilitako metodo ez-gainbegiratuei esker espazio N-dimensional hauetan hainbat propietate aljebraiko betetzen direla bermatzen da (Gittens *et al.*, 2017). Esanahi antzeko hitzak beren artean gertu eta esanahi desberdina dutenak urruti egon ez ezik, hitzen arteko eragiketa aljebraiko oinarrizkoek ere antzekotasunaren nozioari erantzuten diete. Hainbat hitz-bektore desberdinen baturaren emaitzak batu diren hitzen errepresentazioen konbinazioa islatzen du. Ondorioz, hitzen kontzeptuen errepresentazio abstraktuekin lan egiteko bereziki aproposak dira hitz-bektoreak. Gainera, ikasketa-metodoak berak morfologia, sintaxia edo semantikarekin lotutako teknika inpliziturik erabili ez arren, sortutako bektoreetan horien inguruko ezagutza kodetuta islatzen da. Horregatik, hitz-bektoreak ataza sintaktiko zein semantikoetan ebaluatzean emaitza onak eskuratzen dira, adibidez, analogia datu-multzoetan. Analogia-atazetan hitzen arteko antzekotasuna neurtzea da helburu nagusia, eta, horretarako, hitz pareari dagozkion hitz-bektoreak eskuratzen dira eta haien arteko kosinu-distantzia neurtzen da. Bektoreen arteko angeluak hitzen antzekotasuna islatzen du.

Bektoreak corpus handietatik ikasten dira teknika desberdinak erabilita, C-BoW, Skip-gram eta Glove dira gaur egunera arte ezagunenak eta gehien erabilitakoak (ikus 2.2 Irudia). Hirurak hipotesi berdinean oinarritzen badira ere, modu desberdinean erauzten dituzte hitz-bektoreak. Hipotesi horren hitzetan antzeko hitzak antzeko kontestuetan agertzen direnez, teknika horiek corpus handiko agerkidetzetan (agerkidetza-matrizeak) oinarritzen dira. Hitzen agerkidetza-matrize erraldoi hauek hitzak hainbat dimentsio abstraktu erabiliz errepresentatzen dituzte, non dimentsio bakoitzak tasun semantiko edota sintaktiko abstraktu bat islatzen duen. Hitzen agerkidetza-matrizetatik abiatuta posible da neurona-sareetan oinarritutako hizkuntza-ereduak ikastea, **hitzen errepresentazio distribuzionalak**, alegia.

Mikolov *et al.* (2010) autoreek neurona-sareetan oinarritutako hizkuntza-ereduak ikasteko bi proposamen egiten dituzte: bata, C-BoW (Continuous Bag-of-Words), non hitz jakin baten errepresentazio distribuzionala inguruko leiho baten eraginaren ondorioz eguneratzen den; eta, bestea, Skip-gram,



2.2 irudia: Hitz-bektoreak ikasteko erabilitako metodo nagusienetako batzuk: C-BoW eta Skip-gram. Iturria: Euskal Wikipedia, `https://eu.wikipedia.org/wiki/Word2vec`

non, aurreko metodoaren aurka, inguruko leihoan dauden hitzen errepresentazio distribuzionalak eguneratzen diren uneko hitz jakin baten eraginez. Glovek, aldiz, oso modu desberdinean erauzten ditu hitzen errepresentazio distribuzionalak. Lehenik, corpus osoa irakurtzen du eta agerkidetza-matrizea osatzen du, eta behin hori egindakoan, optimizazio funtzio batzuen bitartez hitz-bektoreak erauzten ditu zuzenean (Pennington *et al.*, 2014).

Kontzeptuen errepresentazio distribuzionalak ikasteko arloa hain da zabala, ezen hitzen hitz-bektoreak ez ezik, **karaktereen hitz-bektoreak** ikasteko ere erabili direla. Lan batzuen arabera, esaterako, (Santos eta Zadrozny, 2014) karaktereen hitz-bektoreak morfologia konplexua duten hizkuntzentzat bereziki interesgarriak direla frogatu da. Badirudi, hitz-bektoreek informazio sintaktiko eta semantikoa kodetzen duten heinean, karaktereen hitz-bektoreek informazio morfologikoa modu esanguratsuagoan kodetzen dutela (Kim *et al.*, 2016b).

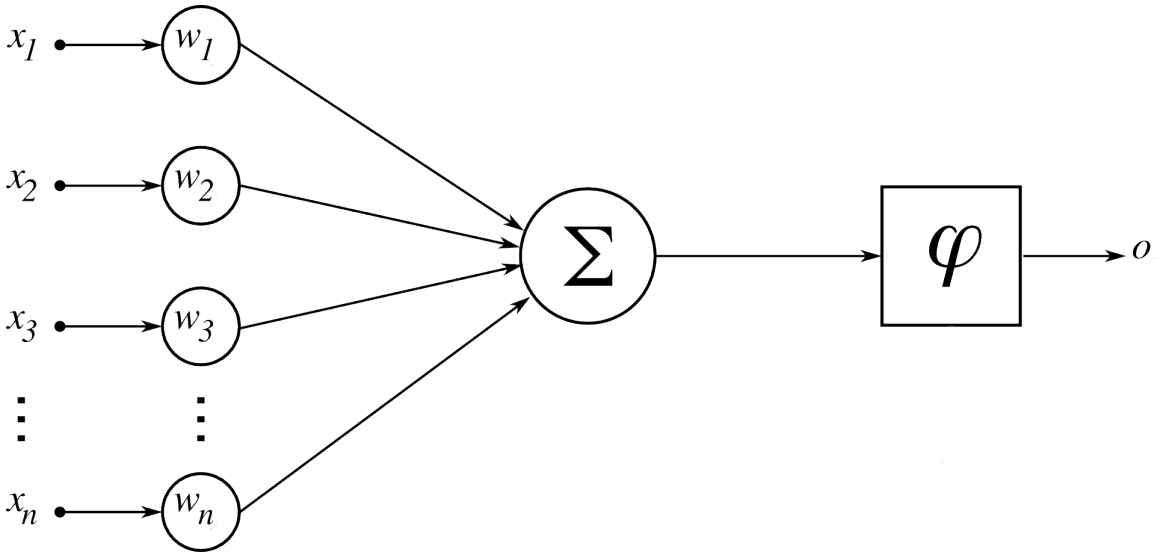
Oro har, semantika distribuzionala arrakastatsua izaten ari bada ere, espazio semantikoetan oinarritutako errepresentazioek badituzte beren mugak eta hainbat ebatzi gabeko arazoei aurre egin beharko zaie etorkizunean. Eztabaida gehien sortu duen arazoetako bat **konposizionaltasuna** da, hau da, hitz-bektoreek hitzetatik haratagoko segmentuen errepresentazioak osatzeko

gaitasuna. Komunitate zientifikoan behin baino gehiagotan zalantzan jarri da ea luzera finkoko eskalarren bektore batek luzera arbitrarioko esaldi baten esanahia islatzeko nahikoa izango ote den ala ez. Itzulpen automatikoko atazetan ezagutza kodetzeko teknika horrek esaldi luzeetara eskalatzeko arazoak agerian jarri baditu ere, momentuz arazoari aurre egiteko proposamenak egin dira, hain zuzen ere, **atentzio-mekanismoak** (Vaswani *et al.*, 2017).

Atentzio-mekanismoak sarrerako bi tentsore irteerako eskalar batekin mapatzen dituzten funtzioak dira. Eskalar honek sarrerako bi tentsoreen arteko osagaien batura eskalatua islatzen du, hau da, tentsoreen interakzio-maila kodetzen eta kuantifikatzen du. Era horretan, hizkuntzaren prozesamenduaren alorrean hitzen arteko interakzioa modelatzeko erabili izan dira atentzio-mekanismoak, askotan hitz-bektoreen arteko biderketa eskalar gisa egikarituak. Atentzio-mekanismoak bereziki itzulpen automatikoan erabili izan dira, metodo horien bitartez itzulpen-sistemek fokoa itzuli behar duten hitz jakinetan jartzeko aukera dutelako, esaldi osoari begiratu beharrean. Esaldi laburren itzulpena egiteko atentzio-mekanismoen beharra hain nabaria ez izan arren, esaldi luzeetarako ezinbestekoa da mekanismo horien erabilera (Bahdanau *et al.*, 2015).

Hitz-bektoreen gaitasunaren beste ahulgune bat lokuzioen edo esamoldeen errepresentazioari dagokio, arlo horretan nahiko mugatuak daudela ondorioztatu baita, baita, polisemiarekin lotutako ebaluazioetan ere (Liu *et al.*, 2015). Hainbat lan egon dira arazo horri aurre egin nahian, ezagunenen artean Upadhyay *et al.* (2017) autoreena, non esanahi anitzeko hitz-bektoreak ikasten dituzten polisemiari aurre egiteko. Hitz-bektoreen beste arazo garrantzitsu bat polaritate negatiboko elementuak behar bezala errepresentatzea da, arazo hau hitz-bektoreak entrenatzeko metodoari lotuta dago gainera. Hitz-bektoreak inguruko kontestua osatzen duen leiho txiki batetik ikasten direla eta, "ondo" edo "gaizki" moduko hitzek errepresentazio oso antzekoak kodetzen dituzte. Ondorioz, hitz-bektoreak polaritatea identifikatzea garrantzitsua den atazetan erabiltzen direnean arazo bat bilakatzen dira (Wang *et al.*, 2015).

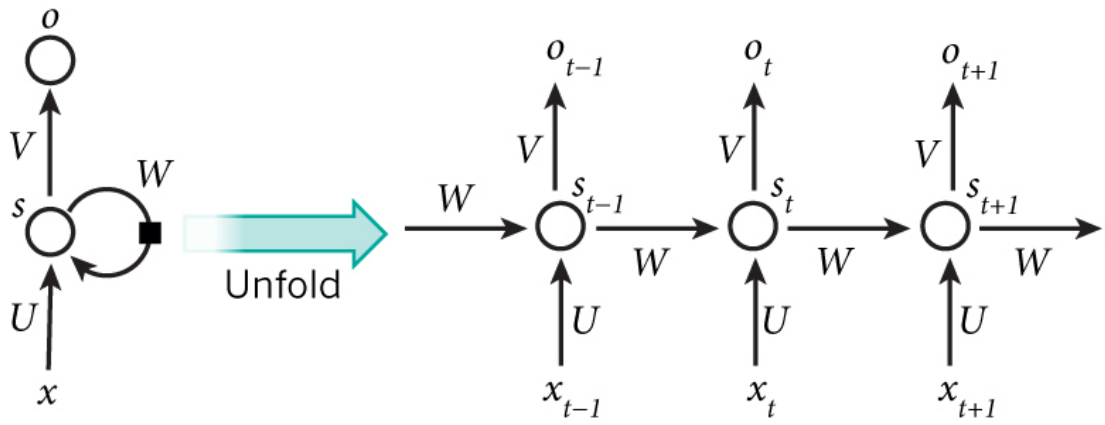
Kontzeptuen errepresentazio distribuzionaletik eskutik helduta aurkitzen ditugun arkitekturak dira **neurona-sareak**. Biak batera konbinazio paregabea



2.3 irudia: Aurrerantz elikatzen den neurona-sare sinple baten diagrama. Iturria: Euskal Wikipedia, `https://eu.wikipedia.org/wiki/Neurona-sare_artifizial`

osatzen dute, neurona-sareek errepresentazio distribuzionaletatik ezaugarri abstraktuak automatikoki erauzi ditzaketelako. Neurona-sareen kateaketek geroz eta abstrakzio maila altuagoko ezaugarriak erauzten doaz, eta, maila altuko atazak modu eraginkorrean ebazteko gai dira. Neurona-sareen motibazio nagusia giza burmuinaren funtzionamendua simulatzea da, non neuronak beste neuronekin elkar eragiten duten estimulu elektrikoen bitartez sinopsiak sortuz. Neurona-sare artifizialak giza burmuineko neurona-sareen simulazio xume bat badira ere (Kriesel, 2007), adimen artifizialeko ataza desberdin askotan aurretik ezagututako ikasketa automatikoko sistema ugari gainditu dituzte (Young *et al.*, 2017).

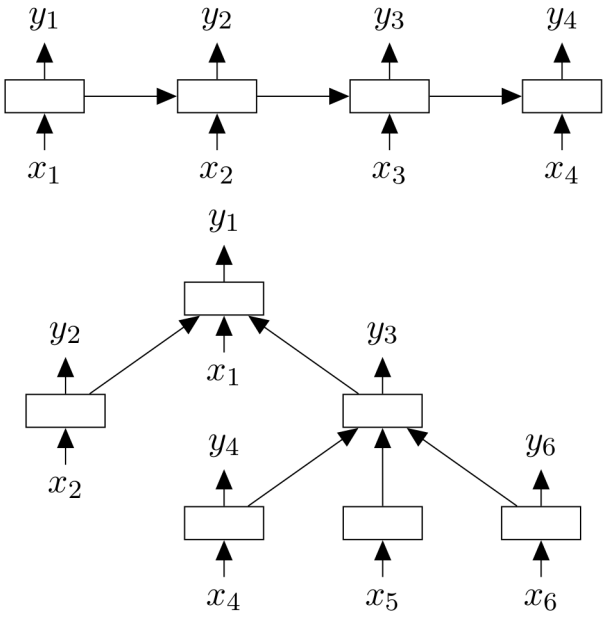
Aurrerantz elikatzen diren neurona-sareak (ikus 2.3 Irudia) dira aurkitu daitezkeen neurona-sare oinarrizkoenak. Sare hauek sarrera bat irteera batekin mapatzeko funtzio arbitrario bat hurbiltzen ikasten dute, atazaren arabera sailkapena edo erregresioa izan daitekeena. Funtzio horren ikasketa neurona-sareari ikasteko adibideak emanez egiten da (Rumelhart *et al.*, 1986), ikasketa-prozesuan neurona-sareak bere arkitektura osatzen duten



2.4 irudia: Neurona-sare errepikakor baten diagrama. Iturria: Wildml neurona-sare tutoriala, `http://www.wildml.com`

neuronen pisuak doitzen doa, eta, era horretan, helburu funtzio bat optimizatzen du. Sare mota horretan –izenak dioen moduan– ez daude atzerantz egiten duten konexiorik, ez eta konexio errekurtsiborik. Neuronen arteko konexio guztiek aurrerantz egin behar dute beti, eta konexio kopuruaren arabera osoak edo partzialak izan daitezke. Sare horiek arkitektura konplexuagoak osatzeko erabiltzen dira, kasu: arkitektura errepikakorrak, errekurtsiboak eta konboluzioan oinarritutakoak.

Sare errepikakorrak (ikus 2.4 Irudia) sekuentzia bat prozesatzeko diseinatuak daude (Elman, 1990). Oinarrizko neurona-sare baten erabilpen errekurtsiboan oinarritzen dira, non urrats bakoitzean sortzen duten irteera hurrengo urratserako sarrera bilakatzen den. Denboran zehar informazioa mantentzeko abilezia atxikitzen zaie, eta, horretarako, urratsez urrats eguneratzen doan memoria bat mantentzen dute. Memoria horren inguruan egiten duten kudeaketaren arabera sare errepikakor desberdinak aurkitu daitezke, ezagunenen artean: arruntak (*Vanilla networks*), LSTM sareak (*Long Short-Term Memory*) eta GRU (*Gated Recurrent Unit*) sareak. Oinarrizko sare errepikakorrek urrats bakoitzean filosofia berdinari jarraituz uneko memoria eguneratzen duten bitartean, LSTM eta GRU neuronak erabiltzen dituzten sare errepikakor aurreratuagoek ate bidezko mekanismoak erabiltzen dituzte

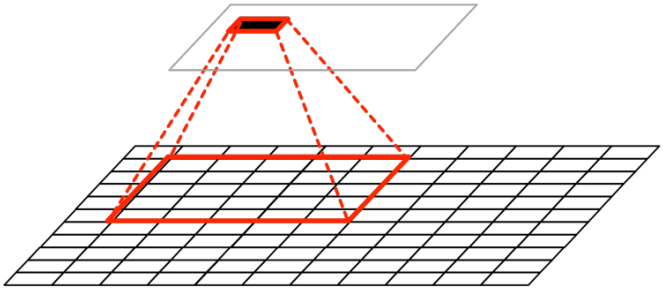


2.5 irudia: Goian neurona-sare errepikakor baten diagrama eta behean neurona-sare errekurtsibo baten diagrama. Iturria: (Socher *et al.*, 2013).

eguneraketa horiek kontrolatzeko. Modu horretan, urrats bakoitzean memoriaren gainean idazketak edo irakurketak egitea ahalbidetzen edo saihesten dute (Hochreiter eta Schmidhuber, 1997a), eta dependentziak askoz modu optimoagoan kudeatzeko gai direla erakutsi da (Chung *et al.*, 2014). Gainera, neurona-sareekin lotutako hainbat arazo[1] ekiditen dituztela ere frogatu da (Bengio *et al.*, 1994). Sare horiek bereziki aproposak dira HPko atazak ebazteko, hizkuntza hainbat hitzen sekuentzia gisa adieraztea tribiala delako (Sutskever *et al.*, 2014).

Sare errekurtsiboak (ikus 2.5 Irudia) sare errepikakorrak orokortzeko ahaleginaren ondorio dira, hizkuntzen egitura sintaktiko adarkatua ustiatzea helburu dutenak. Oinarrian, sare errepikakorrak linealak diren heinean, sare errekurtsiboak zuhaitz-egitura konplexuak erabili ditzaketen arkitekturak

[1] *Exploding* edo *Vanishing gradient* arazoa, sareak atzeranzko propagazio metodoarekin doitzerakoan agertzen dena.



2.6 irudia: Konboluzio neurona-sare baten diagrama. Iturria: (Gu *et al.*, 2017).

dira. Tai *et al.* (2015) autoreen esanetan, hizkuntzek dituzten egitura errekurtsiboak direla eta, hitzak eta segmentuak sintagmetan egituratu ditzaketen egitura hierarkikoak erabiltzea egokiagoa da arkitektura linealak erabiltzea baino. Egitura errekurtsiboa duten neurona-sare hauetan elementu ez-terminalen errepresentazioa bere azpinodoen errepresentazioaren konbinazioak emana da (Socher *et al.*, 2013).

Konboluzio-sareek matematikako konboluzio-eragileari zor diote izena. Konboluzio batean filtro jakin bat seinale baten gainean aplikatzen da, horrela, konboluzioan oinarritutako neurona-sare batean sarrerari neurona-sare bat –filtroa– aplikatzen zaio irteera gisa aktibazio mapa edo ezaugarri mapa bat lortuz (LeCun *et al.*, 1989) (ikus 2.6 Irudia). Neurona-sare guztietan gertatzen den moduan, sareak berak doitzen ditu filtroaren parametroak, eta, era horretan, konboluzio-sareak berak ezaugarri interesgarriak erauzten ikasten duela esan daiteke. Konboluzio-eragileak normalean laginketa-eragile baten aurretik erabili ohi dira, horrela, pausoz pauso, sarreraren dimentsioa murriztuz doa etengabe ezaugarri interesgarrienak azaleratzen diren bitartean. Konboluzio-sareek arrakasta handia izan dute konputagailu bidezko ikusmen-atazetan (LeCun *et al.*, 2015), eta berehala hedatu dira beste arloetara. HPn ez dute horren arrakasta handia izan, baina neurona-sare errepikakorren mailako emaitzak lortu dituzte alor askotan hitz-multzoen inguruko ezaugarriak modu oso eraginkorrean erauzteko abilezia dutelako (Young *et al.*, 2017).

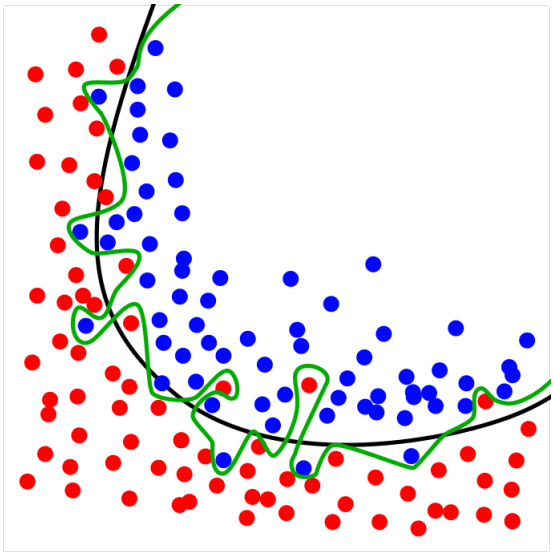
Konboluzioan eta errepikakortasunean oinarritutako neurona-sareen norgehia-

goka garbia dago, ondorio garbirik ez ordea: atazaren arabera batak edo besteak funtzionamendu optimoagoa izan baitezake. Sare errepikakorrek ez bezala, konboluzioak ezin ditu luzera begirako dependentziak modu optimoan mantendu, ez eta hitzen jatorrizko sekuentzia, aitzitik, sarreratik ezaugarri garrantzitsuak erauzteko abilezia handia atxikitu zaie. Kontrara, sare errepikakorrek informazio guztia kodetzen dute beren memorian, baita horren interesgarriak ez diren kontzeptuak ere, horregatik, informazio ez erabilgarri asko kodetzen dutela eztabaidatu izan da (He eta Lin, 2016; Yin *et al.*, 2016).

Neurona-sareen mugak

Neurona-sareen arrakastari kontrajartzen diren bi arazo aipatuko ditugu azpiatal honetan: garatzen diren neurona-sareen arkitekturen konplexutasuna eta interpretagarritasuna. **Sistemen konplexutasunari** dagokionez, geroz eta neurona-sareen kateaketa gehiago erabiltzen dituzten sistemak garatzeko joera nabarmendu da ikerlarien artean (Prijatelj *et al.*, 2017), aitzitik, sistema sakon horiek sortzen dituzten arazoak ere aski ezagunak dira: *overfitting* gisa ezagutzen den terminoa, gehiegi ikastea, alegia (ikus 2.7 Irudia). Gehiegi ikastea arazo larria da adimen artifizialaren arloan sistema batek datu-multzo zehatz batean oso emaitza onak izatera eramaten baitu; baina bere orokortzeko ahalmena erabat mugatzen du, sistemaren benetako erabilgarritasuna nulua izatera pasatuz.

Overfittinga ekiditeko proposamen desberdinak daude: bata, neurona-sarea erregularizatuko duen metodo bat erabiltzea, kasu: dropout; eta, bestea, sistema sinpleagoak eta orokortzeko gaitasun hobea dutenak sortzea. Neurona-sareak sarrera modelatzen duten funtzio matematikoak ikasteko erraztasun handia duten sistemak direnez (Cybenko, 1989), kateatzen diren neurona kopuruaren arabera gehiegi ikasteko arazoa areagotzen da, eta, ondorioz, sisteman overfittinga izateko aukera handitzen da. Hain zuzen ere, dropout metodoa (ikus 2.8 Irudia) neuronak sistematik isolatzeko –ez entrenatzeko– teknika bat da, oso emaitza onak erakutsi dituenak. Etorkizunean sistemen konplexutasuna neurririk gabe inkrementatzeak ekar litzakeen arazoak aurreikusiz, aspalditik ikerlarien artean emaitza onak eskuratu ditzaketen sistema sinpleak hobesten dira, konplexuen aurka. Alor horri dagokionez ez-

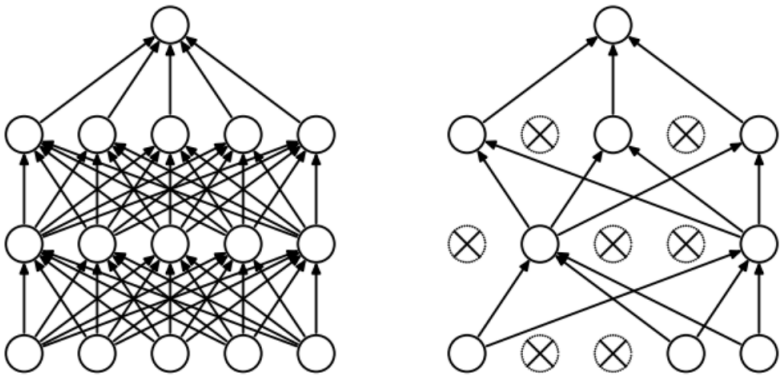


2.7 irudia: Gehiegi doitutako diskriminatzaile bat orokortzeko abilezia hobea duen diskriminatzaile baten ondoan. Iturria: Ingelesezko Wikipedia, `https://en.wikipedia.org/wiki/Overfitting`

tabaida handia dago aurrekarietan (Ba eta Caruana, 2014), eta neurona-sare konplexuen beharra mugatzeko proposamen desberdinak egin dira neurona-sare sinpleak erabilita (Hinton *et al.*, 2015).

Tesi honetan sistema sinple eta malguak sortzeko joerari eutsiko diogu antzekotasuna eta inferentzia egiteko sistemak garatzean. Sistema sinpleak erregularizatzeko eta optimizatzeko metodo egokien aurrean, sistema konplexuen pareko emaitzak lortu ditzakete (Eigen *et al.*, 2013), entrenatzeko behar duten denbora askoz murritzagoa izanik. Gainera, oro har, sistema sinpleak interpretagarriagoak izateko joera dute, erantzuteko azkarragoak eta entrenatzeko arinagoak. Gainera, egitura gutxiago inposatzen dutenez orokortzeko gaitasun altuagoa ere izan ohi dute beste domeinu batzuetara mugitzerakoan.

Bigarren arazo nagusia **neurona-sareak ulertzeko** zailtasunak dira. Hitzbektoreen dimentsioak inplizituki abstraktuak direnez, bereziki zaila egiten da espazio semantiko distribuzionaletan oinarritzen diren sistemak ulertzea. Gure ustez, adimen artifizialaren arrakasta adimen artifiziala bera ulertzeko



2.8 irudia: Dropout teknikak neuronak isolatzen ditu entrenamenduan, parametroak mugarik gabe doitzeko aukerak minimizatuz. Iturria: "Dropout in (Deep) Machine learning", `https://medium.com/@amarbudhiraja`

gizakiok dugun gaitasunarekin lotzen da. Azken finean, adimen artifiziala gizakion mailako gaitasuna duten atazak burutzeko nahi dugu, baina adimen artifizialeko metodoek hartzen dituzten erabakiak eta jokaerak ulertzeko gai ez bagara, nekez ulertuko dugu zergatik egiten duten egiten dutena.

Neurona-sareetan oinarritutako sistemek ikasten dituzten errepresentazioak interpretatzeko hainbat lan egin dira. Lan gehienak **neurona-sareen gaineko ulermen-azterketa hau** irteera geruzen aktibazioekin erlazionatzen saiatu dira (Towell eta Shavlik, 1993), adibidez, irteerak ulertzeko erregelak sortuz (Fu, 1994). Lan horiek hainbat aurrerapauso izan badituzte ere, neurona-sareen benetako ahalmena sareko ezkutuko geruzetan dago (Olah *et al.*, 2018), errepresentazio abstraktuak erauzten ikasten dituzten geruzetan, alegia. Ezkutuko geruzak interpretatzeko oso lan gutxi burutu dira, besteak beste, Murdoch *et al.* (2018) autoreena. Lan horri esker neurona-sareen tarteko errepresentazio abstraktuak bi faktore nagusitan deskonposatzen dira aljebra linealeko metodoen bitartez: uneko hitzek inposatzen duten aldaketan eta kontestuak inposatzen duen aldaketan. Teknika horiei esker sentimendua aztertzea garrantzitsua den atazetan ulermen-maila altuagoa izatea lortu da.

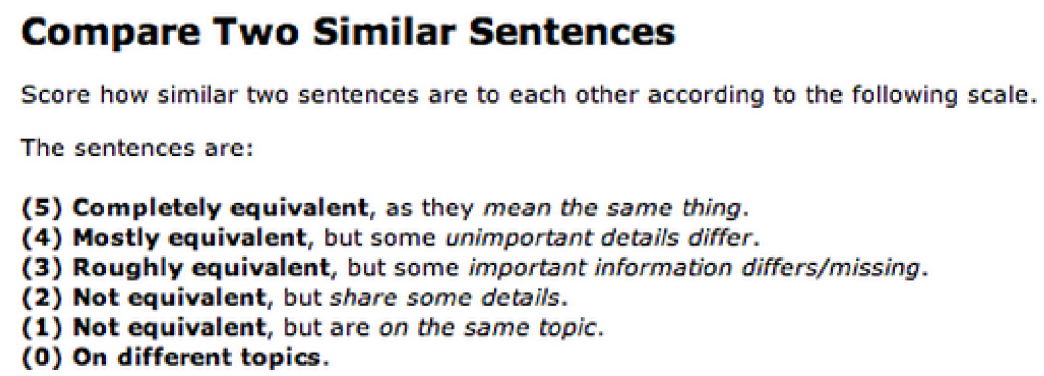
Ulermen-maila areagotzeko ataza bat plazaratzen dugu tesi honetan, **iSTS**,

atazan sistemek antzekotasun semantikoa eta inferentzia logikoa bere horretan erabili ez ezik, hartutako erabakiak maila finean adierazi behar dituzte; beren erabakiak azalduz. Maila fineko lerrokatze hauei esker sistema adimendunek modu abstraktuan hartzen dituzten erabakiak hobeto ulertzea lortzen da, eta, ondorioz, sistemen HU maila areagotzea. Hitzen arteko lerrokatzea aspalditik ezaguna bada ere (Sultan *et al.*, 2014), tesi honek lehen aldiz hitzak baino luzeagoak diren **n-grama arbitrarioak lerrokatzeko**, eta lerrokatzeak **antzekotasun eta logika balioekin aberasteko** proposamenak biltzen ditu. Gure ustez, iSTS sistema konplexuen **benetako hizkuntza-ulermenaren maila** kuantifikatzen laguntzen duen ataza da.

2.3 Esaldi-mailako ebaluazioa

Esaldi-mailako antzekotasun semantikoa eta inferentzia logikoa HUren alorreko bi ataza garrantzitsuenetarikoak dira. Ataza horietan hainbat sistema ebaluatu dira denboran zehar sistemen errepresentazio abstraktuen eta ulermen-mailaren kalitatea neurtzeko. Azken urteetan oinarrizko ataza horiek indarberritzen ari dira hizkuntzalaritza konputazionalean eredu desberdinen hizkuntza-ulermenaren maila ebaluatzea bereziki garrantzitsua delako. Izan ere, maila altuko ataza askoren errendimendua azpitik lan egiten duten sistema adimendunen hizkuntza-ulermenaren gaitasunak emana da, adibidez: konbertsaziorako agenteena. Arrazoi horiek direla eta azkenaldian antzekotasun semantikoko eta inferentzia logikoko datu-multzo berriak plazaratzeko eta sistemak ebaluatzeko joera nabarmendu da.

Esaldi-mailako antzekotasun semantikoak (*Semantic Textual Similarity* edo STS) esaldi pare baten arteko lotura semantikoa kuantifikatzea du helburu, horretarako eskala kuantitatibo bat erabiliz (Agirre *et al.*, 2012). Esaldi pareari emandako balio horrek parearen arteko lotura semantikoa islatzen du, erabateko baliokidetasun semantikotik baliokidetasun ezara (ikus 2.9 Irudia). Gaur egun, STSren aplikazio-esparrua oso zabala da baliokidetasuna islatzen duen balio hau hainbat ataza desberdinetarako oso erabilgarria baita ezaugarri gisa. STSren aplikazio-esparrua hain da zabala, ezen irakaskuntzaren alorrean ikasleen erantzunak kalifikatzeko ere erabili dela (Sultan *et al.*, 2016). Lan horretan autoreek ikasle baten erantzuna eta erreferentziaren ar-

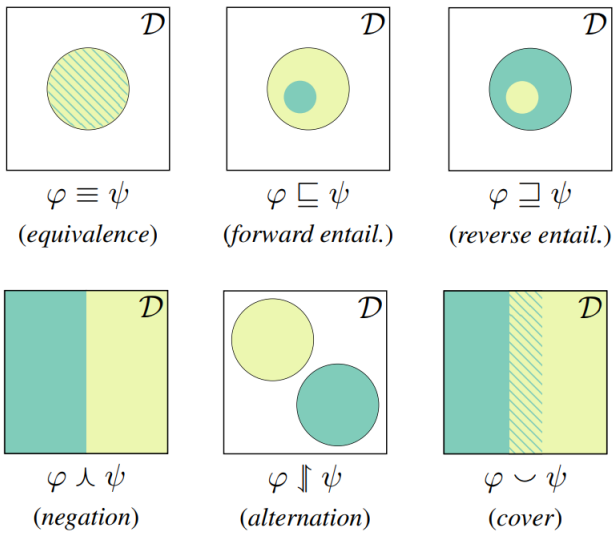


2.9 irudia: Esaldi pareen antzekotasun semantikoa kuantifikatzeko eskala. Iturria: (Agirre *et al.*, 2012).

teko STS balioa erabiltzen dute ikaslearen kalifikazioa esleitzeko. Hala ere, hurbilpen horrek hainbat arazo ditu ikaslea ebaluatzeko, ikaslearen erantzunak eta erreferentziak hitz komunen bat izatea baino metodo osoagoak eta interpretagarriagoak behar direlako.

Esaldi-mailako antzekotasun semantikoak inferentzia logikoarekin erlazio handia duen ataza da, baina inferentzia logikoak direkzionala den kategoria bat esleitzen duen heinean –lotura logikoa dagoen ala ez–, antzekotasun semantikoak graduala eta bidirekzionala den balio kuantitatibo bat esleitzen dio bikoteari. Inferentzia logikoak (*Natural Language Inference* edo NLI) kontzeptuen arteko logika mailako ondorioak egitea du helburu (ikus 2.10 Irudia). Munduko kontzeptuen arteko inferentziak egitea behar-beharrezko ataza da adimen artifizialaren alorrean, eta HPn ez ezik ikusmen eta robotikan ere inferentzia logikoan oinarritutako sistemak oso erabiliak dira, eta, horregatik, aspalditik interes handia piztu duen alorra da (Angeli eta Manning, 2014). Inferentzia logikoan testu batek (T) hipotesia (H) ondorioztatzen du baldin eta soilik baldin testutik hipotesia inferitu –ondorioztatu– badaiteke (T → H).

Esaldien arteko antzekotasun semantikoa eta inferentzia logikoa neurtzea HPrako eta, bereziki, HUrako garrantzitsuak dira aplikazio-esparru zabala dutelako. 2.11 irudian oinarrizko ataza bakoitzeko adibide bana ikus daiteke.



2.10 irudia: Esaldi pareen erlazio logikoa zehazteko balio kualitatiboak. Iturria: (Angeli eta Manning, 2014).

Neurona-sareen arrakastaren ondorioz metodo horiek **ebaluatzeko** proposamen eraginkorren beharra areagotu da. Oro har, bi ebaluazio mota egin izan dira sistema horien kalitatea neurtzeko: **barne-ebaluazioak** eta **kanpo-ebaluazioak**. Barne-ebaluazioetan neurona-sareko helburu funtzioa optimizatu dugun ataza berdinean ebaluatzen dugu sistema, aitzitik, kanpo-ebaluazioetan sistema bere atazatik kanpo ebaluatzen dugu. Azken horietan sistema osoaren zati bat ebaluatu dezakegu, adibidez, neurona-sare konplexuen lehen geruzak ebaluatzea ohikoa da beren errepresentazio-maila atazatik kanpo aztertzeko; edota, sistema osoa guztiz desberdina den ataza batean ebaluatu dezakegu (*transfer-learning*) nahiz eta azken kasu horretan sistema birfindu izan ohi da normalean (Kaiser *et al.*, 2017). Hizkuntza-ulermenaren alorrean geroz eta ohikoagoa da bi motako ebaluazioak burutzea sistemetan, esaterako: Jernite *et al.* (2017); Conneau *et al.* (2017) lanetan autoreek hizkuntza-ulermenerako sistema burutu nahi duten atazan ebaluatzeaz gain, sistemen azpiko geruzen kanpo-ebaluazioa egiten dute analogia estiloko datumultzoetan beren sistemak bereganatu duten hizkuntza-ulermenaren ideia zehatzagoa izateko. Kanpo-ebaluazio hau burutzeko metodo gainbegiratuak

```
Antzekotasun semantikoa
=======================
A turtle walks over the ground.
A large turtle crawls in the grass.
Antzekotasun-balioa: 3.75

Inferentzia logikoa
===================
A white dog is chasing a stuffed animal.
The animal is sleeping.
Logika-erlazioa : Kontraesana
```

2.11 irudia: Antzekotasun semantikoari eta inferentzia logikoari dagokien adibide bana.

eta ez-gainbegiratuak erabiltzen dituzte. Metodo gainbegiratuak ez bezala, metodo ez-gainbegiratuen algoritmo sinpleek, kasu, kosinu-distantziak, benetan jatorrizko sistemak bereganatu duen hizkuntza-ulermenaren maila interpretatzeko bidea errazten du –algoritmoaren sinpletasuna dela eta–.

Analogiaren edo antzekotasunaren kanpo-ebaluazio hauetan, sintaxiaren zein semantikaren ulermen-maila ebaluatzea da helburua. Sintaxia ebaluatzen denean sistemak sortu dituen errepresentazio abstraktuak ebaluatzen dira galdera baten aurrean, eta itzultzen duen emaitzaren zuzentasuna neurtzen da, adibidez: "Zein da aditz jakin baten iragana?" edo antzekoak. Semantika ebaluatzen denean, ordea, hitzen arteko erlazioen hurbiltasuna neurtzen da, esaterako, mota honetako galderekin: "ardoa eta mahatsaren erlazio bera mantentzen duen hitza itzuli garagardoa emanik". Mota horretako ebaluazioez gain hitz, sintagma edo esaldien antzekotasuna ere erabili ohi da sistema konplexuen errepresentazio abstraktuek barneratzen duten hizkuntza-ulermenaren maila ebaluatzeko (Agirre *et al.*, 2009).

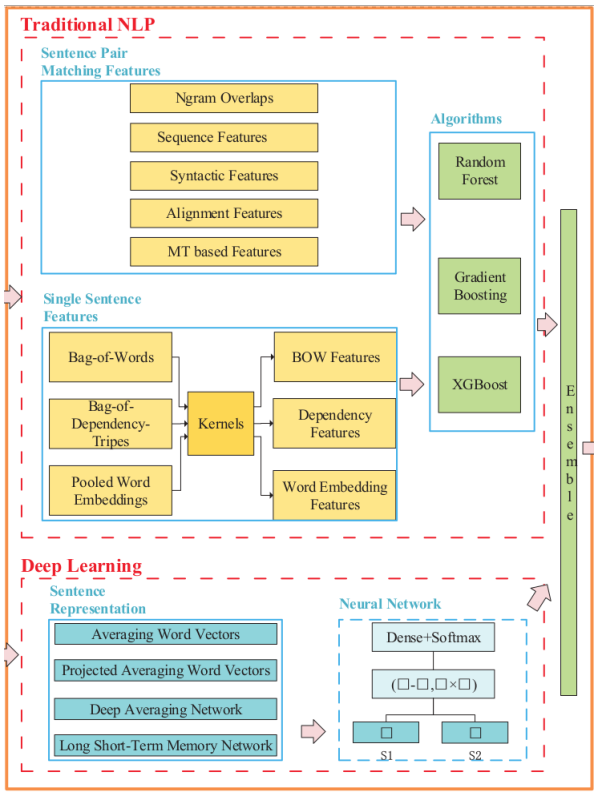
2.4 Esaldi-mailako sistemak

Neurona-sareen arrakastaren ondorioz zientzialariek berehala diseinatu dituzte STSn zein NLIn emaitza onak lortzeko gai diren neurona-sareetan oinarritutako sistemak. Jarraian **emaitzarik onenak** dituzten sistema batzuk deskribatuko ditugu.

ECNU (Tian *et al.*, 2017). Antzekotasun semantikoari dagokion atazetan oso emaitza onak lortzen dituen eta neurona-sareetan oinarrituta dagoen sistema-multzoa da ECNU. Hizkuntza bakar batean edo hizkuntza anitzen artean esaldi pareen artean antzekotasun semantikoa kalkulatzeko gai da sistema hau, honetarako, Googlen itzultzaile automatikoa erabiltzen du hizkuntza desberdinen artean pibotatzeko.

Arkitekturari dagokionez, bukaerako STS balioa hainbat sistemaren bozketaren ondorioz lortzen da, kasu: ezaugarri-zerrendetan trebaturiko erregresioa egiten duten ikasketa automatikoko sistemak; eta hainbat neurona-sare. 2.12. irudian ECNUren arkitektura nagusia ikus daiteke. Irudian ikusten den moduan arkitektura hiru modulu nagusitan banatuta dago: HP modulua, neurona-sare modulua eta sistema-multzo modulua. HP moduluak jatorrizko esaldien ezaugarri sintaktiko eta semantikoak erauzten ditu, besteak beste, esaldien arteko n-gramen gainezarpena, sekuentzia-ezaugarriak, ezaugarri sintaktikoak, lerrokatze gainezarpenak, eta itzulpen automatiko metrikekin lotutako ezaugarriak. Neurona-sare modulua jatorrizko esaldien hitzak bektore bakar batera kodetzeko erabiltzen da, eta, ondoren, STS balioa bektore horren arteko antzekotasunetik erauzten da neurona-sare desberdinak erabilita (LSTM sareak, sare siamdarrak eta DA sareak). Azkenik, sistema-multzo moduluak aurreko bi moduluetatik lortutako emaitzen bozketa egiteaz arduratzen da, bukaerako STS balioa lortuz.

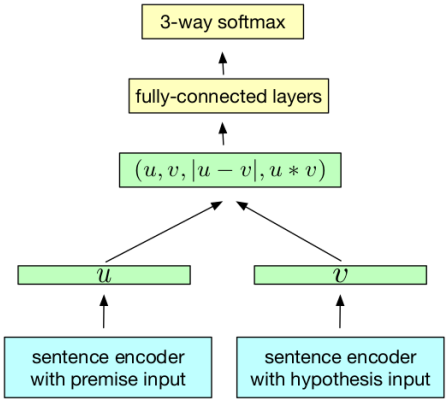
Sistema horrek SemEval atazako emaitzarik onenetarikoak lortu ditu 73.16 pearson punturekin ingelesezko STS 2017 datu-multzoan. Emaitza hori atazako baselinea baino 20 pearson puntu hobea da, eta bigarren sailkatutako parte-hartzailea baino 5 pearson puntu hobea.



2.12 irudia: ECNU arkitekturaren diagrama. Iturria: (Tian *et al.*, 2017).

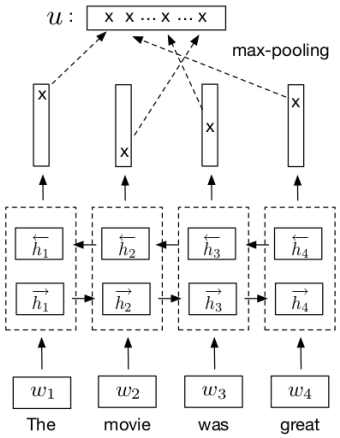
Bi-LSTM-Max + AIINLI (Conneau *et al.*, 2017). Esaldien errepresentazio unibertsalak kodetzeko enkoderrak trebatzen dituen arkitektura bat deskribatzen da lan honetan, antzekotasun semantikorako eta inferentzia logikorako baliagarria dena. Enkoder hauek trebatzeko, arkitekturaren aldetik, neurona-sare errepikakor estandarrak erabiltzen dira (LSTM sareak eta GRU sareak), eta, baita, atentzioan oinarritutako neurona-sareak eta hierarkikoki lan egiten duten CNN errekurtsiboak. Sare errepikakorren kasurako erabiltzen duten enkoderraren eskema 2.13. irudian ikus daiteke.

Enkoderra erabiliz esaldi parearen errepresentazio abstraktuak erauzi ondoren neurona-sare gehigarriak erabiltzen dituzte STSri edo NLIri dagokien azken balioak lortzeko. Balio hau sailkapen edo erregresio bidez erauzten dute NLIren edo STSren kasurako, sare errepikakorren konposizionaltasuna



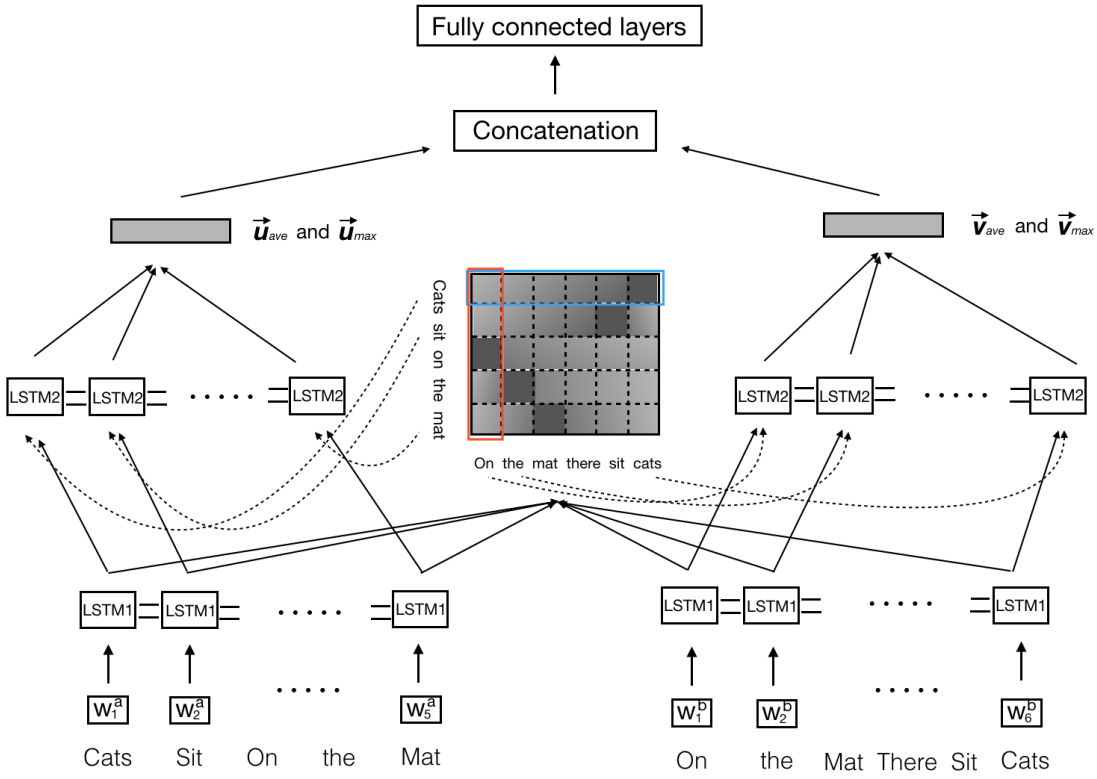
2.13 irudia: Enkoderraren eskema. Iturria: (Conneau *et al.*, 2017).

erabiliz, hurrenez hurren. 2.14. irudian neurona-sare gehigarri hauen eskema ikus daiteke.



2.14 irudia: Konposizionaltasuna lantzeko neurona-sareen eskema. Iturria: (Conneau *et al.*, 2017).

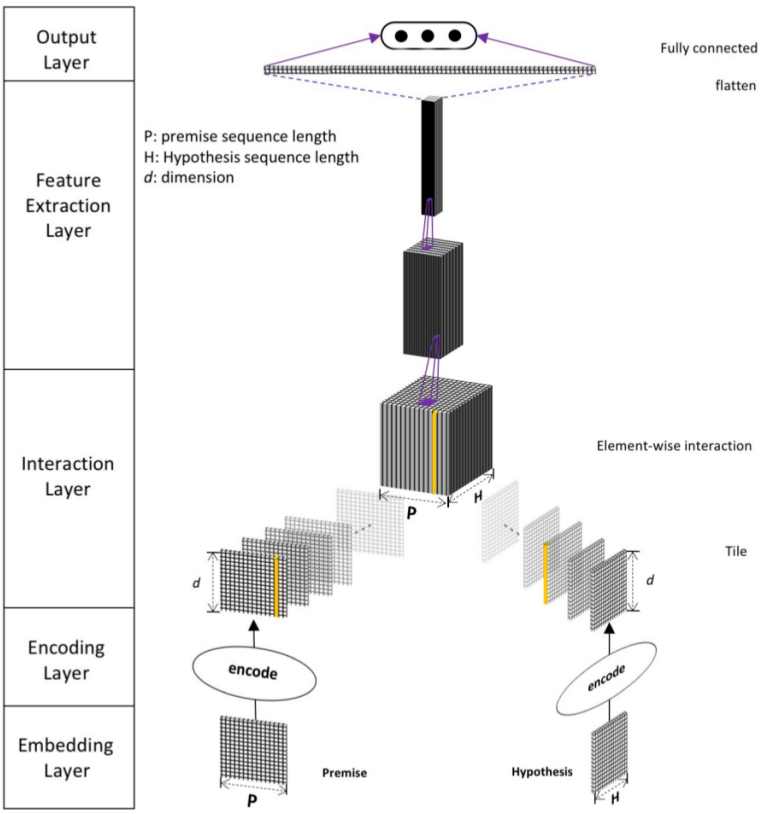
Sistema honen bitartez autoreak hainbat atazatan emaitza onak lortzeko gai dira: 3.4.1. atalean aurkeztuko ditugun SICK datu-multzoaren STS eta NLI adarretan % 86.3ko eta % 88.4ko zehaztasuna lortzeko gai dira, datu-multzo horietan ezagutzen diren emaitza onenetarikoak biak.



2.15 irudia: ESIM arkitekturaren diagrama. iturria: (Lan eta Xu, 2018).

Enhanced Sequential Inference Model (ESIM) (Chen *et al.*, 2016) LSTM sare-kateaketetan oinarritzen den sistema da ESIM, inferentzia logikoak egiteko lehenbizi sekuentzia-kate bat osatzen du esaldien errepresentazio abstraktuak sortzeko eta, azkenik, esaldien errepresentazio horiek alderatzen ditu sailkapena egiteko gai den neurona-sare gehigarri bat erabiliz. 2.15. irudian sistema honen diagrama ikus daiteke.

Hitzen konposizionaltasuna lantzeko kateatzen dituen LSTM sareen arkitekturaren arabera sistemaren bertsio desberdinak ezagutzen dira: $ESIM_{seq}$ (LSTM neurona-sare errepikakor sekuentzialak erabiltzen dituena), $ESIM_{tree}$ (LSTM neurona-sare errekurtsiboak erabiltzen dituena) eta $ESIM_{seq+tree}$ (Aurreko bien konbinazioa). Gainera, sistema horrek hitz bakanen gaine-



2.16 irudia: DINN arkitekturaren diagrama. Iturria: (Gong *et al.*, 2017).

ko lerrokatze- edo atentzio-mekanismo bat ere erabiltzen du, 2.15. irudiaren erdian ikus daitekeen moduan. Mekanismo horri esker, ESIM esaldi pareko hitzen arteko lerrokatzeen garrantzia neurtzeko gai da, eta garrantziaren arabera hitzen pisuak eguneratzeko.

ESIM sistemak oso emaitza onak lortu ditu NLI motako atazetan, bereziki 3.4.1. atalean aurkeztuko ditugun Multi NLI eta SNLI datu-multzoetan % 72.3ko eta % 86.7ko zehaztasuna lortzeko gai da, hurrenez hurren. Ezagutzen diren emaitza onenetatik oso gertu biak.

Densely Interactive Inference Network (DINN) (Gong *et al.*, 2017)

Karaktere-bektoreak, hitz-bektoreak eta ezaugarri-zerrendak erabiltzen di-

tuen sistema-multzo errekurtsiboa da. Sistemaren geruzek hierarkia bat osatzen dute, non geruza bakoitzak jatorrizko esaldien espazio semantikotik ezaugarri desberdinak erauzteko ardura duen. DINN sistemaren arkitektura 2.16. irudian ikus daiteke, eta bost modulu nagusiz osatua dago.

Lehenik, hitz-bektore geruza aurkitzen da, geruza horretan sistemak hitz bakoitzari dagokion hitz-bektorea esleitzen dio. Gainera, hainbat aurreprozesu ere egikaritzen dira, besteak beste, entitate izenak erauztea, informazio sintaktikoa erauztea edota korreferentziak ebaztea. Bigarren geruza enkoderrari dagokio, bertan neurona-sare errepikakorrak edo errekurtsiboak erabiltzen dira hitzen konposizionaltasuna lantzeko. Geruza horrek hitzen arteko dependentziak modelatzea du helburu, hitzak errepresentatzeko erabiltzen diren bektoreak testuinguruarekin hornituz. Sistemaren hirugarren geruza interakzio-geruza da, eta geruza horretan hitz pare guztientzat interakzio-tentsore bat sortzen da. Atentzio-mekanismo tradizionalen kontrara, sistema horrek emaitza gisa tentsore bat –eta ez eskalar bat– sortzen duen atentzio-mekanismo berritzaile bat erabiltzen du, autoreen esanetan hitzen arteko interakzioak hobeto modelatzeko gai dena. Geruza horren azken lana interakzio-tentsore horiek konexio ahulen eta zuzenen bitartez esaldien errepresentazio abstraktuekin konbinatzea da. Hitz pare guztientzat eratutako tentsoreekin DINN sistemak kubo bat sortzen du, hitz-mailako interakzio guztiak jasotzen dituena. Interakzio-kubo hau erabiliz laugarren geruzak ezaugarririk garrantzitsuenak erauzten ditu, garrantzirik gabeko ezaugarriak baztertzeko eta interesgarrienak azaleratzeko helburuarekin. Ezaugarriak erauzteko laugarren geruza hau ikusmen-atazetan motibatutako sistemez baliatzen da lana egikaritzeko, konboluzio-sareetan oinarritutako neurona-sareetan, alegia. Azkenik, bosgarren geruzak, aurrerantz elikatzen den neurona-sare bat erabiltzen du bukaerako inferentzia-kategoria itzultzeko.

ESIM sistemaren antzera, DINN sistemak ere NLI datu-multzoetan emaitza onak lortzeko gai da. Zehazki Multi NLI datu-multzoan % 80ko zehaztasuna lortzeko gai izan da, azken hilabeteetan aurrekarietan ezagutu den emaitzarik onena (ESIM baino 8 puntu gehiago). SNLI datu-multzoari dagokionez ESIM sistemaren pareko emaitzak lortu ditu: % 88.9 (ESIM baino 2 puntu gehiago).

```
Water is split, providing a source of electrons and protons
(hydrogen ions, H+) and giving off O2 as a by-product. Light absorbed by
chlorophyll drives a transfer of the electrons and hydrogen ions from water
to an acceptor called NADP+

1) What can the splitting of water lead to?
  A) Light absorption
  B) Transfer of ions
```

2.17 irudia: ProcessBank datu-multzoko adibide bat. Bertan biologiako prozesu baten deskribapena, galdera eta erantzun posibleak ikus daitezke.

2.5 Ebaluazioa irakaskuntzaren alorrean

Irakaskuntzaren alorreko HU sistemak ebaluatzeko datu-multzoak aztertuko ditugu atal honetan. Horretarako, galdera laburrei emandako erantzunak automatikoki ebaluatzeko atazak aztertuko ditugu, besteak beste, ProcessBank ataza (Berant *et al.*, 2014), MCTest ataza (Richardson *et al.*, 2013) eta Semeval 2013ko 7. ataza (Dzikovska *et al.*, 2016). Aipatutako lan horietan autoreek hezkuntzaren alorrari begira kontribuzio handiak egin zituzten. Ekarpen nagusienen artean ikasketa automatikoan oinarritutako sistemak entrenatzeko baliabideak eta gidalerroak plazaratzea dira bereziki aipagarriak.

Irakaskuntzaren alorrean aztertuko dugun lehen baliabidea da **ProcessBank**. Biologiako prozesu konplexuak modelatzen saiatzen den iturria da (Berant *et al.*, 2014). Paragrafo bat, galdera bat eta bi erantzun agertzen dira deskribatzen den biologiako prozesu bakoitzeko, eta, guztira, 200 prozesu biologiko deskribatzen dira. Prozesuan deskribatzen diren gertaeren eta entitateen artean inferentziak egitea beharrezkoa da erantzun egokia hautemateko. 2.17 irudian adibide bat ikus daiteke.

Gure helburua esaldi-mailako HU lantzea denez gero, etorkizuneko lanetarako uzten dugu baliabide horren analisi sakonagoa, gainera, zalantzazkoak dira datu-multzoaren gainean neurona-sareetan oinarritutako sistemek lortuko lituzketen emaitzak, baliabideak adibide kopuru urria baitu neurona-sareak entrenatzeko eta optimizatzeko. Hala eta guztiz ere, zehaztasunari eta konplexutasunari dagokionez, irakaskuntzaren domeinuan egin den saia-

kera anbiziosoenetarikoa dugu ProcessBank, inferentzia-gaitasun oso altua eskatzen baitu.

```
James the Turtle was always getting in trouble. Sometimes he'd reach into
the freezer and empty out all the food. Other times he'd sled on the deck
and get a splinter. His aunt Jane tried as hard as she could to keep him
out of trouble, but he was sneaky and got into lots of trouble behind her
back.

One day, James thought he would go into town and see what kind of trouble
he could get into. He went to the grocery store and pulled all the pudding
off the shelves and ate two jars. Then he walked to the fast food restaurant
and ordered 15 bags of fries. He didn't pay, and instead headed home.

His aunt was waiting for him in his room. She told James that she loved him,
but he would have to start acting like a well-behaved turtle.After about a
month, and after getting into lots of trouble, James finally made up his
mind to be a better turtle.

1) What is the name of the trouble making turtle?
    A) Fries
    B) Pudding
    C) James
    D) Jane
2) What did James pull off of the shelves in the grocery store?
    A) pudding
    B) fries
    C) food
    D) splinters
3) Where did James go after he went to the grocery store?
    A) his deck
    B) his freezer
    C) a fast food restaurant
    D) his room
4) What did James do after he ordered the fries?
    A) went to the grocery store
    B) went home without paying
    C) ate them
    D) made up his mind to be a better turtle
```

2.18 irudia: MCTest datu-multzoko adibide bat. Bertan anotatzaileek asmatutako ipuina, galderak eta erantzun posibleak ikus daitezke.

Irakaskuntzaren alorrean HU ebaluatzeko beste baliabide bat dugu **MCTest**. MCTest (Richardson *et al.*, 2013) zazpi urteko haurrentzat 660 istorio biltzen dituen baliabidea da. Istorioak anotatzaileak asmatuak dira, eta, hauen helburua sistema desberdinen hizkuntza-ulermenaren maila ebaluatzea da, sistemek istorioen inguruan idatzitako galderak erantzun behar baitituzte. Hain zuzen ere, asmatutako ipuin bakoitzeko lau aukera anitzeko galdera

daude, 2.18 irudian ikus daitekeen moduan. Helburua sistemek ipuina irakurri ondoren galdera bakoitzari dagokion erantzun zuzena hautatzean datza. MCTest moduko datu-multzoetan ondo aritzeko gai liratekeen sistema adimendunak tutore-sistemetan integratu zitezkeen lehen hezkuntzako irakurmenarekin lotutako lanak automatikoki zuzentzeko.

Baliabide hau ere oso interesgarria izan arren, bi arrazoi nagusi daude erabiliko ez dugula azaltzeko: alde batetik, MCTesten erantzunak dokumentumailan erantzun behar direlako, eta, tesi honetan, esaldi-mailako HU dugulako helburu, eta, bestetik, istorioak asmatutako ipuin independenteak direnez, bereziki zaila egiten delako erantzun zuzena zentzuz inferitzeko behar den ezagutza sistema adimendunei helaraztea. Hau da, iturri ezagunetatik –espazio semantikoetatik, ikus 2.2 atala– lortu daitekeen munduaren ezagutza nahikoa ez delako ataza horretarako. Hori dela eta, etorkizuneko lan gisa uzten dugu baliabide honen inguruko azterketa sakonagoa. Are gehiago, burutu genuen azterketaren arabera neurona-sareetan oinarritutako sistemek, kasu Kapashi eta Shah (2015) autoreena, leiho mugikor sinple batean oinarritutako sistema batek baino emaitza okerragoak lortzen zituztela ikusi genuen.

Zientzia desberdinen domeinuan ikasleek galdera irekiei emandako erantzun libreak ebaluatzeko saiakera dugu **SemEval 2013ko 7. ataza** (Dzikovska *et al.*, 2016). Irakaskuntzako sistema adimendunei begira ikasleen erantzunak ulertzea eta bakoitzari dagokion kalifikazioa esleitzea[2] ziren ataza honetan egin beharreko lanak. Lan horrekin batera bi datu-multzo plazaratu ziren: Beetle and ScientsBank, eta, horien bitartez, autoreek irakaskuntzaren alorreko HParen teknologia sustatzea zuten helburu. Alde batetik, Beetle corpusa oinarrizko elektrizitate eta elektronikaren inguruan bildutako ikasleen erantzunez osatua dago (Dzikovska *et al.*, 2010b), eta, beste aldetik, ScientsBank corpusa zientziaren esparruko hamasei domeinutako interakzioz osatua dago (Nielsen *et al.*, 2008). Datu-multzo horiek hainbat galderaz, erreferentzia-erantzunez eta ikasleek laborategietan emandako erantzunez osaturik daude, 2.19 irudian datu-multzo hauen inguruko adibide bana

[2]Kategoria posibleen artean: zuzena, erdizka edo hein handi batean osatu gabea, kontraesana, garrantzirik gabekoa eta domeinuz kanpokoa.

```
The sand and flour in the gray material from mock rocks is separated by
mixing with water and allowing the mixture to settle.
Explain why the sand and flour separate

Reference Answers:
  A) The sand particles are larger and settle first.
  B) The flour particles are smaller and therefore settle more slowly
Student responses:
  A) The sand and flour separate because sand floats to the top and the
  flour stays on the bottom (contradictory)
  B) Because sand is heavier than flour (partially correct incomplete)


Explain why you got a voltage reading of 1.5 for terminal 1 and the positive
terminal

Reference Answer:
  A) Terminal 1 and the positive terminal are separated by the gap
Student responses:
  A) Because there was not direct connection between the positive terminal
  and bulb terminal 1 (correct)
  B) Voltage is the difference between a positive and negative end on a
  battery. (irrelevant)
  C) Tell me the answer (out of domain)
```

2.19 irudia: SemEval 2013ko 7. atazako adibide parea, ScientsBank eta Beetle corpusetatik erauzita hurrenez hurren. Bertan galdera, erreferentziazko erantzunak eta ikasleen erantzunak ikus daitezke, dagokien kalifikazioekin.

irakur daitezke.

Parte-hartzaileak ebaluatzeko hiru eszenario desberdin erabili ziren, ikasketarako erabilitako datuen erlazioaren arabera: ikusi gabeko erantzunak, ikasketarako datu-multzoan ikusitako galderen erantzun berriez osatua; ikusi gabeko galderak, ikasketarako datu-multzoan ikusi gabeko galdera eta erantzunez osatua, baina ikusitakoen domeinua mantentzen zutenak; eta, azkenik, ikusi gabeko domeinuak, guztiz berriak diren galdera eta erantzunez osatua. Horretaz gain, antolatzaileek ebaluaziorako neurri desberdinak ere erabili zituzten aipatutako eszenario bakoitzean parte-hartzaileen zuzentasuna neurtzeko. Emaitzak agerian utzi moduan, garbi geratu zen zorroztasuna igo ahala parte-hartzaileen emaitzak okertzen zirela (Dzikovska *et al.*, 2013), eta sistemen ulermen-maila areagotzea beharrezkoa izango zela eszenario errealista batean orokortzeko gaitasun altua lortu nahi bada behintzat. Tesi honetan

autoreek plazaratutako Beetle datu-multzoa berrerabiliko dugu ataza berri bat sortzeko, iSTS (Lopez-Gazpio *et al.*, 2016a), esaldi barruko kontzeptuen ulermen-maila areagotzea helburu izango duena.

Datu-multzo hauekin ikertzaileek kontribuzio berritzaileak egin zituzten eta interes zein parte-hartze handia piztu zuten irakaskuntzaren aplikazioen alorrean. Gainera, atazetan parte hartzen duten sistema desberdinak ebaluatzeko irizpide komun bat ezartzea ere lortu zuten. **Irakaskuntzako HU sistemei dagokionez** helburu nagusia ikasketa automatikoan oinarritutako sistema malguak sortzea da, erregeletan oinarritutako sistemen ahuleziak gaindituko dituztenak. Horretarako, bi proposamen nagusitu dira:

1. Logikan oinarritutako metodoez baliatuta erantzunaren eta erreferentziaren artean ondorioztatu daitezkeen proposizio kopuruarekiko kalifikazioa esleitzea.

2. Erantzuna erreferentziarekin alderatzea hurbiltasun sintaktikoaren eta semantikoaren araberako kalifikazioa esleituz. Kasu horietan kalifikazioa erantzuna eta erreferentziaren arteko hitzen gertutasunak markatzen du.

Inferentzia logikoan oinarritutako metodoek mugak izan ohi dituzte kontzeptuak behe-mailan identifikatzeko, eta kontzeptuen arteko erlazioak zehaztasunez hautemateko (Levy *et al.*, 2013), bi arrazoi nagusi direla eta (Dzikovska *et al.*, 2013): alde batetik, logikaren alorreko loturaren nozioa eta irakaskuntzaren alorreko adierazpen nozioa –kontzeptu jakin bat barneratu dela jakinaraztea– desberdinak direlako, eta, bestetik, metodo hauek terminologiarekin oso zorrotzak diren heinean, ikasleek erantzunak idaztean terminoak ekiditeko edo eraldatzeko joera dutelako.

Bestalde, **hurbiltasun semantikoan oinarritutako metodoak** modu orokorrean lan egiteko diseinatu direnez malguagoak dira, baina, aitzitik, ez dira gai esaldiko hitz gakoen arteko erlazio logikoak modu finean hautemateko (Mohler *et al.*, 2011). Adibide gisa, logikaren ikuspuntutik auto jakin bat ibilgailua dela ondoriozta daiteke, baina ez ibilgailu bat autoa denik –ibilgailu mota ugari daudelako–. Aitzitik, antzekotasunak autoaren eta ibilgailuaren

arteko hurbiltasun semantikoaren kalkulua ahalbidetuko liguke, baina ez bi kontzeptuen arteko loturaren informaziorik, ez eta esaldietan agertu daitezkeen kontrajarritako kontzeptuen arteko loturen informaziorik ere.

Aipatutako bi teknika nagusi hauek beren aldeko abantailak eta desabantailak dituzte, eta, hauen bitartez hezkuntzaren domeinuan saiakerak egin diren arren teknika mugatuak direla ikusi da, hobekuntzarako tartea daukatenak. Tesi honetan oinarrizko teknika horiek uztartuko ditugu eta feedback esanguratsua itzultzeko metodo berritzailea dela ikusiko dugu. Era horretan, teknika bakoitzak ematen dizkigun abantailaz baliatu gaitezke ahalik eta esaldiaren errepresentazio osoagoak lortzeko.

3. KAPITULUA

Hitz n-gramen arteko atentzio-ereduak

Iñigo Lopez-Gazpio, Montse Maritxalar, Mirella Lapata and Eneko Agirre.
Word n-gram attention models for sentence similarity and inference
Preprint submitted to Knowledge-Based Systems.
ISSN: 0950-7051

Hitzen n-grama arbitrarioak modelatzeko eta lerrokatzeko zehaztapenak proposatzen ditugu lan honetan. Horretarako, neurona-sare arkitektura desberdinek egitura islatzeko dituzten gaitasunak eztabaidatzen ditugu eta gure proposamenak antzekotasun semantikoan eta inferentzia logikoan oinarritutako hainbat datu-multzotan ebaluatzen ditugu. Artikulu honek tesiaren 3. kapitulua osatzen du, eta bai antzekotasun semantikoaren inguruko lanen, baita inferentzia logikoaren inguruko lanen erreferentzia nagusia da.

Word n-gram attention models for sentence similarity and inference

I. Lopez-Gazpio[a], M. Maritxalar[a], M.Lapata[b], E.agirre[a]

[a] *IXA NLP group, Computer Science faculty, University of the Basque Country (UPV/EHU), Manuel Lardizabal 1, 20018, Donostia, Basque Country*
[b] *Institute for Language, Cognition and Computation, School of Informatics, University of Edinburgh 10 Crichton Street, Edinburgh EH8 9AB*

Abstract
Semantic textual similarity and natural language inference are two popular natural language understanding tasks used to benchmark sentence representation models where two sentences are paired. in such tasks sentences are represented as bag of words, sequences, trees or convolutions, but the attention model is based on word pairs. in this article we introduce the use of word n-grams in the attention model. our results on five datasets show an error reduction of up to 41% with respect to the word-based attention model. the improvements are especially relevant with low data regimes and, in the case of natural language inference, on the recently released hard subset of natural language inference datasets.

3.1 Introduction

A major challenge in Computational Linguistics is that of building meaning representation models to enable Natural Language Understanding (NLU). In order to train and evaluate those models the community has proposed several challenges and associated datasets, including Machine Comprehension (MC) (Rajpurkar *et al.*, 2016), Question Answering (QA) (Yang *et al.*, 2015), Automatic Short Answer Grading (ASAG) (Burrows *et al.*, 2015), Natural Language Inference (NLI) (Bowman *et al.*, 2015) and Semantic Textual Similarity (STS) (Agirre *et al.*, 2012). In those tasks, the NLU system needs to pair two text snippets and then provide an output such as the relevance between a question and a text passage (MC), a question and an answer (QA),

the two responses from a teacher and from a student (ASAG), the entailment relation between text and hypothesis (NLI) or the similarity score between two sentences (STS), respectively. In this paper we will focus on the latter two tasks, even the technique can be easily applied to the other tasks.

Computational linguists have used several approaches in the past, with deep learning systems getting consistently the best results when training data is available (Williams *et al.*, 2018; Cer *et al.*, 2017). These systems encode each of the input sentences into a vector using different methods, ranging from simple bag-of-words (BoW) (Parikh *et al.*, 2016), convolutional neural networks (CNN) (Yin *et al.*, 2016), recurrent neural nets such as LSTM (Nangia *et al.*, 2017) to recursive tree LSTM (Tai *et al.*, 2015). Some systems compare the vectors of the input sentences directly, and compute the output without access to the underlying information (Tai *et al.*, 2015; Choi *et al.*, 2017). The most successful systems, though, take also into account word alignment information, usually in the form of word attention models (Parikh *et al.*, 2016; Chen *et al.*, 2016; Gong *et al.*, 2017). Those attention models capture the correspondences between words in the pair of sentences. On the other hand, Lopez-Gazpio *et al.* (2017) observe that alignments between linguistically motivated chunks[1] are very useful in order to capture the semantic relations between two sentences, in the framework of a shared task called Interpretable STS. Despite this observation, alignment and attention models continue to be limited to words.

This article proposes to extend the alignment information from pairs of words to pairs of word n-grams, motivated by the observation of Lopez-Gazpio *et al.* (2017). The use of word n-grams is common practice in statistical language models (Stolcke, 2002). More recently, sentence embedding models have complemented unigram (word) embeddings with bigram embeddings (Pagliardini *et al.*, 2018). In our proposal we model attention as a weight for each possible word n-gram pair [2] instead of each possible word pair. We first extract sequences of contiguous words ranging from one single word to a maximum of N words for both sentence pairs, and build an attention matrix

[1]Chunks are similar to phrases, but do not require full parsing (Abney, 1991).

[2]For the sake of clarity we will use n-gram to mean word n-gram (as opposed to character n-gram) throughout this article.

for all such n-gram pairs. In this work we use recurrent neural networks to represent n-grams, but other options like n-gram embeddings could be used (Zhao *et al.*, 2017).

We explore the effect of the proposed attention model on a competitive BoW system called Decomposable Attention Model (DAM)(Parikh *et al.*, 2016). We show that the n-gram alignment model improves results when compared to DAM with word attention, and that it is a better alternative than modeling context using LSTMs and CNNs. In addition, we train the attention model as a regression module, improving further the results. Our system is evaluated on multiple STS and NLI datasets. It is especially beneficial in datasets with lower amounts of training data and, in the case of NLI, on the hard subset of NLI datasets. Our system also compares well to the state-of-the-art, and shows promise for adding n-gram attention to other systems.

This article is structured as follows. We first lay out the background, including the STS and NLI tasks, followed by the definition of the Decomposable Attention Model. Section 3.3 introduces the proposal to extend the word alignment model. Section 3.4 describes the datasets and results. Section 3.5 presents the comparison to state-of-the-art systems. The final section draws the conclusions and mentions future work.

3.2 Background

In this section we review the the two sentence pairing tasks where we apply the proposed attention model, STS and NLI. In addition, we present the system which we will extend with our N-gram attention model.

STS and NLI

STS (Agirre *et al.*, 2012) aims to measure the degree of semantic equivalence among two textual sentences. STS datasets are composed of input sentence pairs alongside their gold standard scores. Figure 3.1 shows a couple of examples extracted from two distinct STS sources: STS Benchmark (Cer *et al.*, 2017) and SICK textual similarity (Marelli *et al.*, 2014). We review the cited datasets in further detail in Section 3.4.1. The gold standard scores are

obtained by averaging the scores of several annotators, and ranges between 0 and 5. The highest value is for full semantic equivalence and the lowest value for no relation at all.

```
Example 1
=========
A turtle walks over the ground.
A large turtle crawls in the grass.
Similarity score: 3.75

Example 2
=========
The children of a family are playing and waiting.
An Asian man is dancing and three kids are looking.
Similarity score: 1.9
```

Figure 3.1: Examples from Semantic Textual Similarity datasets. See text for further details.

NLI datasets also comprise an input sentence pair and a manually assigned relation label, where the label establishes the entailment relation between the two sentences, which is usually one of entailment, neutral or contradiction (Bowman *et al.*, 2015). NLI is also known as Textual Entailment (*TE*). Figure 3.2 shows three examples extracted from three distinct NLI sources: SICK Textual Entailment (Marelli *et al.*, 2014), Stanford Natural Language Inference (Bowman *et al.*, 2015) and Multi-Genre Natural Language Inference (Nangia *et al.*, 2017). We also review the cited datasets in further detail in section 3.4.1. The first sentence in the pair is said to be the text (T) and the second sentence the hypothesis (H). The annotators need to decide whether T entails the hypothesis H, that is, whether reading T suggests that H is true (Dagan *et al.*, 2006). If not, they need to decide whether T contradicts H. The remaining label is neutral, that is, T neither entails nor contradicts H.

Both STS and NLI are popular evaluation scenarios for semantic representation models, as similarity and entailment relations often involve complex linguistic phenomena. In fact, White *et al.* (2017) have converted several linguistically annotated datasets into entailment pairs. STS and SNLI datasets thus make it easy to judge the degree to which semantic representation models are able to effectively capture some aspects of the meaning of language.

```
Example 1
=========
Sentence 1 : A tiger cub is playing with a ball.
Sentence 2 : A baby is playing with a doll.
Relation label : Neutral

Example 2
=========
A white dog is chasing a stuffed animal.
Sentence 2: The animal is sleeping.
Relation label : Contradiction

Example 3
=========
Sentence 1: Please renew your commitment today.
Sentence 2: A renewal of commitment is required today.
Relation label : Entailment
```

Figure 3.2: Examples from Natural Language Inference datasets. See text for further details.

STS is related to NLI, as argued in (Agirre *et al.*, 2012). They both aim at capturing semantic relationships between the input sentence pairs. STS is symmetric and graded, while NLI is directional and categorical. They each are able to evaluate different traits of semantics, but both include desired requisites for any NLU system. An interesting example of the difference between similarity and inference is to consider the case between pairs of objects that hold the hypernym relation, e.g. two pairs like wildcat-cat and cat-animal. STS defines a similarity value for the pairs, higher for wildcat-cat than for cat-animal, but the same values as for the inverse pairs cat-wildcat and animal-cat. Inference is directional, and thus it captures entailment for wildcat-cat and neutrality for the inverse cat-wildcat, but does not differentiate the different strength of the association in wildcat-cat and animal-cat.

The Decomposable Attention Model (DAM)

There is a growing number of systems pushing the state-of-the-art results on STS and NLI upwards. In this work, we chose to add our n-gram attention model to the Decomposable Attention Model (Parikh *et al.*, 2016) because of its simplicity, low number of parameters and high performance. DAM relies in the key concept that long sentences tend to be complex in structure,

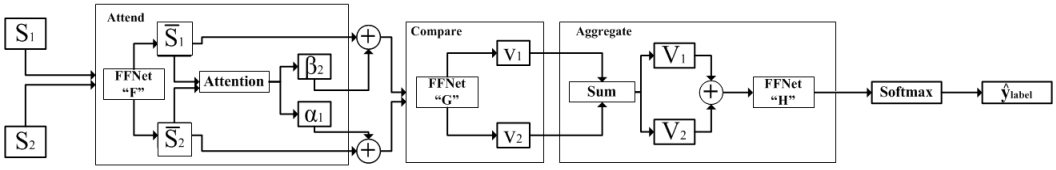


Figure 3.3: Architecture of the Decomposable Attention Model. The figure shows the concatenation of the three main layers of the model: the attention layer, the comparison layer and the aggregation layer. FFNet denotes a feed-forward neural network and the + operator denotes concatenation of vectors.

and, therefore, it is hard for computational models to construct a compact and reliable fixed-size representation that captures the entire meaning of the input. Furthermore, Parikh *et al.* (2016) state that most of the times the alignment among small parts of the content words can lead to successful entailment judgments. Although the system was originally designed for NLI, it is straightforward to adapt it to produce similarity scores, as we will see in the end of this section.

The architecture of DAM is shown in Figure 3.3. It consists of three feed-forward neural networks (F, G and H) structured in three consecutive layers as follows: the attention layer, the comparison layer and the aggregation layer. Each feed-forward network is composed of a single hidden layer and employ rectified linear units (ReLU) as non-linear functions[3]. Input and output dimensionality for all networks is kept constant and is defined by the global hidden size architectural setting.

The attention layer (*Attend* block of Figure 3.3) is where the soft-alignment between input words happen using a variation of neural attention. Given input sentences S_1 and S_2 represented as 2-dimensional tensors[4], the model first linearly transforms the input sentences applying the F network individ-

[3]Feed-forward networks (FFNet) consist of a total of 3 layers: input, hidden and output. Both hidden and output layers contain trainable parameters and the same non-linearity function (ReLU) after the linear transformation.

[4]The first dimension indexes word S_i from sentence S and the second dimension its corresponding word vector.

ually obtaining $\overline{S}_1$ and $\overline{S}_2$ respectively as output, following these equations: $\overline{S}_1 = F(S_1)$ and $\overline{S}_2 = F(S_2)$. Once the input sentences are transformed, the attention layer computes projection vectors Alpha (α_1) and Beta (β_2) as follows:

$$\beta_{2i} = \sum_{j=1}^{|\overline{S}_2|} \frac{\exp e_{ij}}{\sum_{k=1}^{|\overline{S}_2|} \exp e_{ik}} \overline{S}_{2j} \ , \ i \in |\overline{S}_1|$$

$$\alpha_{1j} = \sum_{i=1}^{|\overline{S}_1|} \frac{\exp e_{ij}}{\sum_{k=1}^{|\overline{S}_1|} \exp e_{kj}} \overline{S}_{1i} \ , \ j \in |\overline{S}_2|$$

where $|\overline{S}|$ denotes sentence length and e_{ij} denotes word to word attention computed as the dot product among normalized word vectors: $e_{ij} = \overline{S}_{1i} \cdot \overline{S}_{2j}$. As a result, β_2 contains the weighted sum of words from $\overline{S}_2$ projected onto the first sentence and α_1 contains the weighted sum of words from $\overline{S}_1$ projected onto the second sentence.

The comparison layer (*Compare* block of Figure 3.3) learns to compare the previously aligned words and projections, producing v_1 and v_2 vectors respectively:

$$v_{1i} = G([\overline{S}_{1i} \ ; \ \beta_{2i}]) \ , \ i \in |\overline{S}_1|$$

$$v_{2j} = G([\overline{S}_{2j} \ ; \ \alpha_{1j}]) \ , \ j \in |\overline{S}_2|$$

where the semicolon operation denotes vector concatenation.

The aggregation layer (*Aggregate* block of Figure 3.3) makes the final judgment based on the representation produced by the previous layers. It initially compacts and flattens the vectors containing the comparisons among words and projections, and obtains the final probability distribution estimates over the labels ($\hat{y}_{label}$) using the H network and softmax estimation. The final inference label (y_{label}) is obtained by picking the most probable class.

$$V_1 = \sum_{i=1}^{|\overline{S}_1|} v_{1i}$$

$$V_2 = \sum_{j=1}^{|\overline{S}_2|} v_{2j}$$

$$\hat{y}_{label} = \text{softmax}(H([V_1 \ ; \ V_2]))$$

$$y_{label} = argmax \ \hat{y}_{label}$$

We refer the reader to (Parikh *et al.*, 2016) for further details on the DAM model. In this work, we re-implement the model to use it as a baseline. We call this baseline model DAM BoW. Our baseline model follows the same training criterion as Parikh *et al.* which uses negative log-likelihood as the loss function to be minimized.

DAM was proposed for NLI tasks. In order to adapt it to TS, we change the final layer so that the model performs regression instead of classification. We use the well-known approach by Tai *et al.* (2015) for this task. Following this work, during training the model predicts TS scores as if it were labels, optimizing the Kullback-Leibler divergence. For the task, the TS scores are converted into probability mass estimates over the discrete labels in the TS range. When testing the model to obtain the final predictions (y_{score}), the following formula is used to reconvert the probability mass estimates over the discrete labels into TS scores in the range $[0, 5]$:

$$y_{score} = r^T \cdot \hat{y}_{label}$$

where r is a row vector containing a value for each discrete number in the TS range.

3.3 Extensions to word alignment

The main contribution of this paper is the addition of n-gram attention to DAM. As one could argue that n-gram attention is merely adding context information into the attention model, we also implemented two extensions to the word attention model which add context information to tokens via recurrence and convolution. These two extensions are baselines which the n-gram attention model should outperform to show its value. We will first introduce these extensions and then present the n-gram attention model. In addition, our model benefits from a trainable attention model, which is presented last.

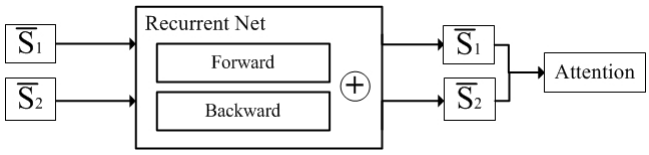


Figure 3.4: Addition to the attend module of DAM to introduce context using recurrence. See text for further details.

Adding context through recurrence

DAM BoW computes context-independent attention scores e_{ij} between words and, after that, re-weights the word vectors of the input sentences using the row-wise or column wise normalized e_{ij} values. As a consequence, the resulting tensors alpha and beta relate to input sentence 2 and input sentence 1 respectively based on e_{ij} values computed out of word to word interaction. In order to extend the word interaction between the input sentences, in this first extension we propose to run a recurrent neural network before the attention mechanism of DAM BoW in order to compute context-based representations of words. The context-dependent representations of words are then used to compute the attention scores e_{ij} in the same manner.

As shown in the schema of figure 3.4 in this extension we propose to modify the representation of every word on $\overline{S}_1$ and $\overline{S}_2$ formalizing it to be the concatenation of the forward and backward output states of a recurrent neural network for that word:

$$\overline{S}_i = [\mathrm{RNN}_i^f(\overline{S}) \ ; \ \mathrm{RNN}_i^b(\overline{S})] \ , \ i \in |\overline{S}|$$

where $\mathrm{RNN}_i^f(\overline{S})$ and $\mathrm{RNN}_i^b(\overline{S})$ denote the output state of the forward and backward passes respectively for word $i \in \overline{S}$. Note that by doing so the dimensionality required to represent each word in the sentence doubles.

Adding context through convolution

As an alternative to recurrence, one can exploit context through convolutions over nearby windows of words. We achieve this by concatenating the feature

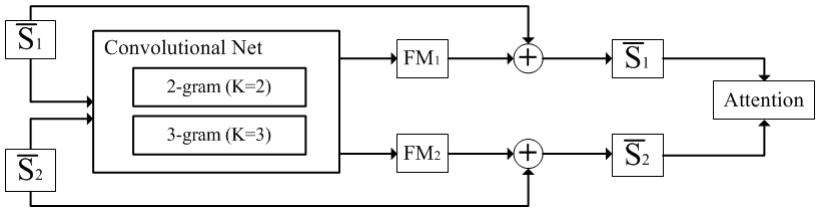


Figure 3.5: Addition to the attend module of DAM to introduce context using convolutions. See text for further details.

maps (FM) learned by convolution filters over input words. In this context, feature maps are defined as:

$$\text{FM} = \text{CNN}(\overline{S}\ ,\ \text{filter size} = K)$$

We tested convolution filters (K) of sizes two and three respectively. A schema showing the changes required for this approach can be seen in Figure 3.5. In a similar way to the extension based on recurrence, this time the dimensionality required to represent each word in the sentence also doubles as we concatenate the previous representation of the word with the learned feature map for every word in the sentence:

$$\overline{S}_i = [\overline{S}_i\ ;\ \text{FM}_i(\overline{S})]\ ,\ i \in |\overline{S}|$$

We apply padding when necessary to maintain the same dimensionality for the input and output vectors.

Word n-gram alignments

As mentioned in the introduction, this paper proposes to replace word alignment by n-gram alignment. Instead of enriching the representations of words using context (as done in the previous subsections), we hypothesize that explicitly representing word n-grams and computing attention between all possible n-gram pairs will perform better. Given the sentence $\overline{S}$ and assuming that $Ngram(\overline{S}, x, y)$ denotes the word n-gram starting from index x up to

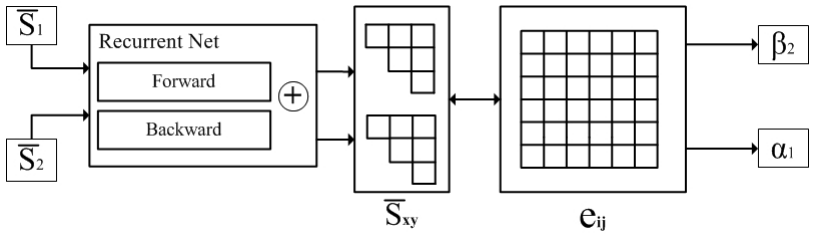


Figure 3.6: Addition to the attend module of DAM to introduce structure using arbitrary n-grams. See text for further details.

y in $\overline{S}$, the representation of the n-gram ($\overline{S}_{xy}$) is obtained as a bi-directional RNN which is run on that sequence as follows:

$$\overline{S}_{xy} = [RNN^f\ (\text{Ngram}(\overline{S}, x, y))\ ;\ RNN^b\ (Ngram(\overline{S}, x, y))\ ]$$

$$1 \leq x \leq |\overline{S}|\ ,\ x \leq y \leq |\overline{S}|$$

The resulting representation is an upper-diagonal matrix composed of all n-grams of $\overline{S}$, where the diagonal represents a 1-gram (single word) and subsequent squares to the right represent longer n-grams, which keep on adding words one by one at a time to the previous n-gram. The maximum size of n-gram to consider is defined by an hyper-parameter (N), $y - x < N$. Thus, the number of n-grams the model handles for sentence $\overline{S}$ is given by $|\overline{S}| \cdot N - \sum_{i=1}^{N-1} i$. When $N = 1$ the number of n-grams is equal to $|\overline{S}|$, that is, the number of elements in the diagonal, and if $N = |\overline{S}|$ the number of n-grams is equal to the number of elements in an upper triangular matrix of size $|\overline{S}|$ which is defined by $\sum_{i=1}^{|\overline{S}|} i$.

Figure 3.6 (middle box) shows the schema for the described architecture to represent n-grams. Given two sentences $\overline{S}_1$ and $\overline{S}_2$ the n-gram attention mechanism defines a matrix (e_{ij}) where i linearizes over the n-grams of $\overline{S}_1$ and j linearizes over the n-grams of $\overline{S}_2$. Figure 3.6 shows the full schema of the DAM N-gram approach. Note that the main difference with regard to DAM BoW resides in that, in this extension, each attention value e_{ij} captures the attention between n-gram i (corresponding to some n-gram $\overline{S}_{xy}$ spanning

from x to y) from sentence $\overline{S}_1$ and n-gram j (corresponding to some n-gram $\overline{S}_{kz}$ spanning from k to z) from sentence $\overline{S}_2$. From another perspective, the attention model linearizes the triangular matrix of possible n-grams, that is, i is the linear index over possible (x,y) tuples and j is the linear index over possible (k,z) tuples.

Attention as an end-to-end trainable module

The usage of distinct attention mechanisms to attend to words has already been explored in the state-of-the art. For instance, Luong *et al.* (2015) define three well-known attention mechanisms. In the cited work the authors consider three distinct alternatives to score the attention between a pair of words: (1) neural attention, which is just the dot product (Bahdanau *et al.*, 2015) $\overline{S}_i \cdot \overline{S}_j$ given $i \in \overline{S}_1$ and $j \in \overline{S}_2$; (2) general attention, which trains a weight matrix implemented as $\overline{S}_i \cdot W_1 \cdot \overline{S}_j$; and, (3) concat attention, which applies a 2-layer transformation to the concatenation of the representation for the words implied in the interaction, implemented as $W_2\tanh(W_1[\overline{S}_i\ ;\ \overline{S}_j])$.

In this work we experiment with all the three variations above, but we adapted the concat attention model to use the same feed-forward neural network with RElUs as in the rest of DAM (FFNet, see Section 3.2.2). We tested all three possibilities (cf. Section 3.4.4). We refer to our implementation of the concat attention as *FF attention*.

$$e_{ij} = \text{FFNet}([\overline{S}_i\ ;\ \overline{S_2}])$$

3.4 Experiments

We now describe the experiments involving the original DAM and the proposed extensions, including the datasets, the evaluation metrics, the experimental setup and implementation details, development experiments and the main results.

Dataset	Train	Dev	Test	Total
STS Benchmark	5749	1500	1379	8628
SICK (TE /TS)	4439	495	4906	9840
SNLI (filtered)	549367	9842	9824	569033
MultiNLI (matched)	392702	9815	9796	412313

Table 3.1: Train, dev and test splits for all five datasets.

Description of the datasets

We evaluated our systems on the most relevant textual similarity and natural language inference datasets. Table 3.1 shows the number of examples for each of the datasets.

Semantic Textual Similarity has been the focus of an annual task until 2017 (Agirre *et al.*, 2012; Cer *et al.*, 2017). STS contributed towards defining an unified framework and stimulate research for evaluating systems that measure the degree of sentence level semantic equivalence. Each year the challenge brought together numerous participants, with new datasets. Recently, the organizers released a dataset that comprises a selection of all datasets, in order to provide a standard benchmark to evaluate different models in a unified framework, the **STS Benchmark** dataset. The selection of datasets includes those in the domain of image captions, news headlines and user forums (see Table 3.2). Note that the development set is partially mismatched regarding the training and test sets. We refer the reader to the official website[5] for further information. An example of this dataset is available in Figure 3.1 (Example 1).

The **SICK** dataset (Marelli *et al.*, 2014)[6], *Sentences Involving Compositional Knowledge*, comprises semantically challenging sentence pairs, which had been semi-automatically selected and manipulated to comprise phenomena such as lexically rich words, contextual synonymy, active and passive changes, syntactic alternations and negation. The dataset was annotated both for textual similarity (**SICK-TS**) and textual entailment (**SICK-TE**).

[5] `http://ixa2.si.ehu.es/stswiki/index.php/STSbenchmark`
[6] `http://clic.cimec.unitn.it/composes/sick.html`

Genre	Train	Dev	Test
Microsoft Research Paraphrase	1000	250	250
SemEval news headlines	1999	250	250
SemEval DEFT news	300	0	0
Microsoft Research video captions	1000	250	250
SemEval Image captions (2014-2015)	1000	250	250
SemEval Image captions (2017)	0	125	125
SemEval DEFT forum crawl	450	0	0
SemEval question-answer pairs in forums	0	375	0
SemEval answer-answer pairs in forums	0	0	254

Table 3.2: Sources used in the STS Benchmark dataset, showing that development set is partially mismatched with regards to the training and test sets.

Sentences come from ImageFlickr[7] and MSR-Video descriptions[8]. More information can be gathered in the official website[9]. We provide two examples of this dataset in figures 3.1 (Example 2) and 3.2 (Example 1), the first annotated with a similarity score and the second with an inference label.

The previous datasets have a relatively small number of training examples. In an effort to mitigate the lack of large-scale resources and scale-up existing resources for machine learning research, Bowman *et al.* (2015) introduced the **Stanford Natural Language Inference** corpus (SNLI[10]). In contrast to previous resources, sentences from SNLI were written by crowd-sourcing in a grounded, naturalistic context, and labels were inferred automatically. Consisting of a total of 570k pairs it is two orders of magnitude larger than all previous resources. Following usual practice, we use the filtered version, where pairs that do not exhibit annotation agreement were removed. An example from this dataset can be read in Figure 3.2 (Example 2).

While SNLI focused on image captions, MultiNLI[11] (Nangia *et al.*, 2017)

[7] `http://nlp.cs.illinois.edu/HockenmaierGroup/data.html`
[8] `http://www.cs.york.ac.uk/semeval-2012/task6/index.php?id=data`
[9] `http://clic.cimec.unitn.it/composes/sick.html`
[10] `https://nlp.stanford.edu/projects/snli/`
[11] `https://www.nyu.edu/projects/bowman/multinli/`

introduces new genres, enlarging the diversity of linguistic phenomena, including temporal reasoning, belief and modality among others. The test subset contains only five of the genres present in train and development. We thus focus on the matched subset of MultiNLI, where the subsets of training and development coming from those five genres are used. An example from the dataset can be observed in Figure 3.2 (Example 3).

Evaluation metrics

Following usual practice we use Pearson product-moment correlation coefficient to report performance on TS datasets and accuracy to report performance on NLI datasets. Pearson measures the linear dependence between a pair of variables and outputs a value in the range $[-1, 1]$. Accuracy states the number of predicted examples that hold the same label with regards to the gold standard annotation divided by the total number of samples in the dataset and outputs a value in the range $[0, 1]$.

Implementation details

We used Pytorch in the implementation. The texts were tokenized and punctuation removed. Regarding hyper-parameters and design options, we run experiments on the development datasets alone.

Following (Parikh *et al.*, 2016) we use pre-trained Glove word embeddings in the input. Glove word embeddings[12] have been broadly used to initialize a wide range of neural network architectures and are based on word co-occurrence counts (Pennington *et al.*, 2014). We tested several versions on development data, with the best results for the embeddings trained on with 840 Billion tokens.

Feed-forward networks used ReLU non-linearity. We tried several approaches for the recurrent neural networks employed in sections 3.3.1 and 3.3.3, including simple Recurrent Neural Networks (RNNs), Gated Recurrent Units (GRUs) (Cho *et al.*, 2014) and Long Short-Term Memory networks (LSTMs) (Hochreiter eta Schmidhuber, 1997b). We empirically found out that LSTMs

[12] `https://nlp.stanford.edu/projects/glove/`

	SICK-TS	STS-B	Multi NLI	SNLI	SICK-TE
Embedding size	300	300	300	300	300
Hidden size	450	450	500	600	1050
Weight decay	5e-5	5e-5	5e-5	5e-5	5e-5
Max grad norm	5.0	5.0	5.0	5.0	5.0
Dropout	0.5	0.5	0.25	0.15	0.15
Param init	1e-2	1e-2	1e-2	1e-2	1e-2
Learning rate	9e-4	1e-4	1e-4	7.5e-5	1.9e-5
Epochs	87	46	98	95	72
Optimizer	Adam	Adam	Adam	Adam	Adam
Max n-gram size	2	2	4	4	4

Table 3.3: Hyper-parameters for the proposed system, DAM N-gram with FF attention. See appendix for the hyper-parameters of the baseline systems.

and GRUs outperform RNNs by large margin, whereas the performance between LSTMs and GRUs was similar, slightly in favor of GRUs. We opted in favor of GRUs as the defined neuron unit is simpler while still keeps the memory gate that makes the difference with respect to standard RNNs. The election of GRUs over LSTMs also favors the time required to train models by large margin. We also follow (Tai *et al.*, 2015) for Textual Similarity tasks so that instead of concatenating word embedding vectors $\vec{A}$ and $\vec{B}$ we concatenate their element-wise difference (distance) defined as $|\vec{A} - \vec{B}|$ and their element-wise product (angle) defined as $\vec{A} \odot \vec{B}$. For NLI tasks, we empirically observed that concatenating $\vec{A}$, $\vec{B}$, $|\vec{A} - \vec{B}|$ and $\vec{A} \odot \vec{B}$ yields slightly better results.

We included dropout in all layers. We noted that high dropout ratios (40% - 50%) are useful in Textual Similarity datasets as the training set is reduced in size, and complex models can easily overfit them. In tasks with larger available resources we did not find dropout to be among the most important hyper-parameters to tune. We tested both Adagrad (Duchi *et al.*, 2010) and Adam (Kingma eta Ba, 2014) optimizers.

We optimize all the hyper-parameters using random search (Bergstra eta Bengio, 2012) which is stated to find better settings in a limited amount of

System	Train	Dev		Train	Dev
DAM BoW	.927	.746	+ FF att.	.946	.765
DAM RNN	.935	.757	+ FF att.	.922	.780
DAM CNN_2	.915	.747			
DAM CNN_3	.971	.774	+ FF att.	.972	.771
DAM N-gram	.930	**.801**	+ FF att.	.928	**.817**

Table 3.4: Development results (Pearson) in the STS Benchmark dataset for distinct approaches. The rightmost columns correspond to the respective DAM versions with FF attention (cf. Section 3.3.4).

time compared to grid search. The hyper-parameters were tuned using the available development set for each dataset separately. The hyper-parameters of our proposed model are described in Table 3.3.

Development of the systems on STS-B

In order to develop the system proposed in Section 3.3, we decided to do some development experiments on the STS Benchmark development dataset first. We chose STS Benchmark because it is smaller than the Natural Language Inference datasets, and, compared to the SICK datasets, it contains a wider range of topics and the development set is partially mismatched regarding the training and test sets (see Table 3.2).

Table 3.4 shows the development results. In the first row we show the DAM architecture (DAM BoW, cf. Section 3.2.2). In the rows below we show the results for the two baseline methods to encode context in the attention model (DAM RNN, cf. Section 3.3.1, and DAM CNN, cf. Section 3.3.2) as well as our proposed model (DAM N-gram, cf. Section 3.3.3). The DAM CNN model includes two rows, as we tested filters of maximum width 2 and 3. The results show that all approaches to encode context improve the results with respect to the original DAM, with RNNs yielding a weak gain, CNNs with width 3 performing better, and with the best results for the n-gram attention model.

The table also shows, in the rightmost columns (denoted by *+ FF att.*), the results when adding the FF attention module (described in Section 3.3.4)

	SICK-TS		STS-B		MultiNLI		SNLI		SICK-TE	
System	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test
DAM BoW	.768	.771	.746	.679	.717	.725	.854	.852	.745	.727
DAM BoW$_{FFatt}$	.802	.794	.765	.726	.681	.676	.855	.854	.765	.766
DAM RNN$_{FFatt}$	.836	.826	.780	.742	.719	.720	.857	.850	.796	.787
DAM CNN$_3$	.811	.814	.774	.741	.722	.721	.852	.856	.789	.781
DAM N-gram$_{FFatt}$	**.860**	**.857**	**.817**	**.773**	**.750**	**.748**	**.867**	**.863**	**.844**	**.840**

Table 3.5: Results for baselines and proposed model in textual similarity (Pearson) and inference datasets (Accuracy). FFatt for FF attention, STS-B for STS-Benchmark.

to the systems in the rows. We report the results for the most significant systems. The FF attention module yields improvements between 2.3 and 1.6 points, except for CNNs, where it does not improve results. We also tested the general attention model (cf. Section 3.3.4), but found that it is below the Feed-Forward attention model by 3 - 1.5 absolute points.

All in all, the best results are obtained by our n-gram attention model with FF attention, with improvements of around 5 points with respect to the original word-based attention model. The results also show that the n-gram attention model is superior to the alternative baselines (RNNs or CNNs) to infuse context information into a word-based attention model. We will confirm these development results when testing on all five datasets.

Main results

We now evaluate the most representative systems on all five test datasets, including textual similarity and inference, as seen in Table 3.5. All hyperparameters were set using the respective development dataset (cf. Section 3.4.3). Regarding the performance of our implementation of DAM (DAM BoW), it is better by around half a point on both MultiNLI and SNLI over the performance of the implementation reported on Gururangan *et al.* (2018), although it is one point below the performance reported by the original authors on SNLI (Parikh *et al.*, 2016).

The table includes also DAM BoW with FF attention (DAM BoW$_{FFatt}$), and

System	MultiNLI hard	SNLI hard
DAM BoW	.563	.712
DAM BoW$_{FFatt}$	.496	.711
DAM RNN$_{FFatt}$	.551	.704
DAM CNN$_3$	.563	.717
DAM N-gram$_{FFatt}$	**.611**	**.734**

Table 3.6: Results (accuracy) for baselines and proposed model in the hard subsets for SNLI and MultiNLI (Gururangan *et al.*, 2018).

the best RNN, CNN and n-gram attention system as reported in the development experiments. The results confirm the trends observed in development: the two baseline methods to encode context in the attention model (DAM RNN, cf. Section 3.3.1, and DAM CNN, cf. Section 3.3.2) improve over the original DAM model, and that both models perform very similarly, although RNNs require the FF attention model to match CNNs. Our proposed model (DAM N-gram, cf. Section 3.3.3) yields the best results in all five datasets. The improvements vary from dataset to dataset, with the biggest gains on SICK-TE (11.3 abs olute points and a relative error reduction of 41%), SICK-TS (8.6 absolute, 38% error reduction) and STS Benchmark (9.4 and 29%, respectively). The gains for the SNLI and MultiNLI datasets are smaller, 1.1 and 2.3 absolute points, 7.4% and 8.4% error reduction, respectively.

We examined the reason for the smaller differences in MultiNLI and SNLI. Gururangan *et al.* (2018) found that significant portions in SNLI (67%) and MultiNLI (53%) could be solved based on the hypothesis text alone, ignoring the premise sentence. This large portion of trivial pairs can make the differences in performance smaller, and they thus released two subsets of MultiNLI and SNLI, the so-called hard subsets. We evaluated our systems on the hard subsets, and found that the ranking of systems does not vary, but the differences in performance are larger (see Table 3.6). The absolute difference between our proposed n-gram attention model with learnable attention and the BoW model is of 2.2 and 4.8 points for SNLI and MultiNLI, respectively, and the error reduction of 7.6% and 11.1%, confirming that the trivial parts of SNLI and MultiNLI dilute performance differences. These results show that our system is specially effective for the more realistic hard pairs of the

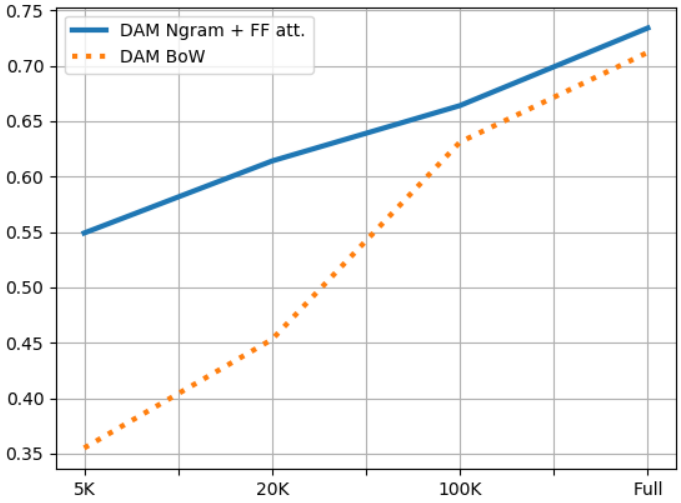


Figure 3.7: Results (accuracy) for different training set sizes in the hard subset for the baseline and proposed model.

NLI datasets.

In addition, we studied whether the amount of training data is an important factor in the performance differences. Figure 3.7 shows the performance on the hard subset of SNLI of relevant systems with smaller subsets of the training data. The figure clearly shows that our proposal is more effective on the smaller subsets. The fact that the performance differences are also larger on the three datasets with smaller amounts of training data (STS-B and the two SICK datasets) seems to confirm that our proposed algorithm is specially effective on low data regimes.

In summary, the results across the five datasets confirm the development results. Our n-gram attention model combined with FF attention is able to provide large performance gains with respect to word-based attention models, including those models using RNNs or CNNs to add context information into the word-based attention model.

3.5 Comparison to the state-of-the-art

Representing and comparing two text snippets as in STS and NLI is a usual benchmark for testing NLU architectures. We will review the most relevant state-of-the-art systems for our work, with emphasis on the attention model that they use. Head-to-head empirical comparison of specific components across architectures is difficult, as the final results of complex systems are affected by several design decisions, including pre-processing, sentence representation, attention model, or final classification/regression layer.

The goal of this section is to show that the performance of the DAM system with our n-gram attention model is competitive with respect to comparable systems, that is, systems which have comparable modules on all layers but attention. We first group the best-known systems, then compare their performance head-to-head in a table, and finally discuss the differences with respect to the two best-performing systems. We classify existing systems according to their representation for the input texts: transfer models, recurrent models, convolutional models and recursive models. Finally, we group ensemble models.

Transfer models employ external learning objectives to train distributed representations of sentences. *Sent2Vec* (Pagliardini *et al.*, 2018), for instance, is an extension of CBOW (Mikolov *et al.*, 2013a) that learns word representations such that each word in the sentence can be predicted based on the average of the representations for the rest of the words in the sentence. Similarity is computed as the cosine between those vectors. To our knowledge it has not been applied to NLI. *SkipThought* (Kiros *et al.*, 2015) is another example of this kind in which the training objective is to maximize the reconstruction of neighboring sentences based on the recurrent representation (using LSTMs) of the current sentence. A classification and regression layer is trained on the respective similarity or NLI dataset, in order to fine-tune the representations to the task. Note that none of these two methods use any attention layer. As an alternative to transfer models, the methods presented below learn the representations directly on the provided training data, with the exception of word embeddings, which are often initialized with pretrained values.

RNN models also encode the meaning of the sentence into a single vector and compute the final label out of it using some kind of recurrent structures such as bidirectional LSTMs. Williams *et al.* (2018) present a baseline system, *Bi-LSTM*, which uses a bidirectional LSTM to encode the meaning of the sentences, and then compute distance and angle vector features between the two sentences, which are fed to a single non-linear layer. This system does not use any attention model. $ESIM_{seq}$ (Chen *et al.*, 2016) is based on bidirectional LSTMS, and introduces a word-based attention layer and an extra layer of bidirectional LSTMs on top of the word-based attention layer. *DINN* (Gong *et al.*, 2017) uses additional features in the input, where each word is represented as a concatenation of a word embedding, character features and syntactic features. In addition, they use multi-head attention, which is an extension of standard word-based attention to a 3D tensor, where the attention between two words is represented as a vector instead of a single scalar. The attention tensor is exploited using deep convolutional networks.

Tree models employ recursive tree-structured neural networks such as Tree-LSTMs to learn to compose the appropriate structure out of the input. *Constituency* and *dependency Tree-LSTM* (Tai *et al.*, 2015) generalize regular linear LSTM chains into Tree-structured LSTM chains. The *Gumbel TreeLSTM* approach (Choi *et al.*, 2017) uses an alternative tree-learning algorithm which dynamically selects candidate nodes using Straight-Through Gumbel-Softmax estimation. The previous two models do not use any attention model.

FF models use feed-forward networks to encode the meaning of the sentence. They include DAM, which uses a word attention model, and therefore our proposal is also a member of this family. More recently, self-attention has emerged as a powerful tool to model intra-sentence dependencies. *Reinforced self-learning* Shen *et al.* (2018) combine soft and hard attention with an emphasis on self-attention and also include multi-head attention. The hard attention module trims the input for the soft attention module, while the soft attention module feeds back signals in the form of rewards to the hard attention module. Both are combined via reinforcement learning. They claim to extract efficiently the sparse dependencies between selected token

System	Type	Attention	MNLI	SNLI	S-TE	S-TS	STSB
DAM BoW	FF	Word	.725	.852	.727	.771	.679
DAM N-gram$_{FFatt}$	FF	N-gram	.748	.863	**.840**	.857	**.773**
Sent2vec	Transfer	-				.620	.755
SkipThought	Transfer	-			.823	.858	
BiLSTM	RNN	-	.669	.815			
ESIM$_{seq}$	RNN	Word	.723[a]	**.867**[a]			
Single DINN	RNN	3D	**.788**	.865[b]			
Constituency Tree-LSTM	Tree	-				.868	.719
Gumbel Tree-LSTM	Tree	-		.860			
Reinforced self-attention	FF	Self		.863		**.872**	
ECNU	Feature	-			.836	.828	
ESIM$_{seq+tree}$	Ensemble	Word		.886			
DINN	Ensemble	3D	.800	.889			
BiLSTM-Max+AIINLI	Ensemble	-			.863	.884	

Table 3.7: Results (accuracy) for our models (first two rows) and representative state-of-the-art models on STS and NLI datasets (see text for references). Best non-ensemble systems in bold, second best underlined. MNLI for MultiNLI, S-TE for SICK-TE, S-TS for SICK-TS and STSB for STS Benchmark. Source of results are the original papers (see text for references), with the following exceptions: [a] (Williams *et al.*, 2018), [b] Gururangan *et al.* (2018).

pairs without involving recurrence or convolutions.

Feature-based models are based on sets of manually designed heuristics encoded as features which are fed to a machine learning algorithm. For instance, *ECNU* (Zhao *et al.*, 2014) combines a total of seventy two features including length differences, word overlap measures weighted with tf.idf, matrix factorization of distributional vectors, overlap of dependencies, antonyms from WordNet, string similarity and co-occurrence-based distributional models. Support vector machines are used to classify (or regress) the target entailment (or similarity) label.

Finally, **ensemble models** obtain improvements combining the output of several models. For instance, *ESIM$_{seq+tree}$* (Chen *et al.*, 2016) combines the recurrent model mentioned above with another system based on Tree-LSTMs.

System	SNLI hard	MultiNLI hard
DAM N-gram$_{FFatt}$	**.734**	.611
ESIM$_{seq}$	.713	.593
Single DINN	.727	**.641**

Table 3.8: Results (accuracy) for proposed model and two competing models in the hard subsets of SNLI and MultiNLI (Gururangan *et al.*, 2018). ESIN and DINN results taken from (Gururangan *et al.*, 2018).

DINN (Gong *et al.*, 2017) does the majority vote of the predictions given by multiple runs of the same model (see Single DINN above) under different random parameter initialization. Alternatively, models which are substantially different can also be combined. *BiLSTM-Max+AIINLI* (Conneau *et al.*, 2017) combines two different recurrent models (LSTMs and GRUs), self-attentive networks and hierarchical convolutional networks.

Table 3.7 shows the results of the systems mentioned above, with Table 3.8 reporting the results on the hard subset of SNLI and MultiNLI. Given that systems have been evaluated in different datasets, the comparison between two systems is limited to common datasets. Ensemble methods, as expected yield the best results in all datasets. We include them for completeness, but we are mainly interested in the comparison between single systems.

The best system in each dataset varies, with one different winner in each dataset, except our proposed system which is the best on SICK-TE and STS Benchmark. Our system performs better than Sent2Vec, Bi-LSTMs, Gumbel Tree-LSTM and ECNU in all datasets in common. The comparison with the rest of the systems is not clear, as our system wins in one dataset but not in the other. The only exceptions are Single DINN, which is better than our system in both MultiNLI and SNLI, and Reinforced self-attention, which is better on SICK-TS and equal on SNLI.

The qualitative comparison between competing systems and ours shows that our n-gram attention model is a module which could be complementary to the components of the other systems and vice-versa. We will now focus on those differences, system by system. For instance, SkipThought trains the LSTM in the input layer on a very large unsupervised task, and reuses it for STS

and NLI. The addition of a n-gram attention model could further improve results, and, on the opposite direction, transfer learning could improve the sentence representations of our system.

In the case of ESIM$_{seq}$, they use Bi-LSTMs both in the input layer and after attention, in the inference layer. This double use of recurrence is complementary to the use of our n-gram attention model, and adding the recurrent networks to our model could further improve results. In any case, the results on the hard subsets (Table 3.8) shows that our system beats ESIM$_{seq}$ on both SNLI and MultiNLI when trivial examples are ruled out.

Regarding Single DINN, it uses a richer input layer, a multi-head attention layer, and convolution and pooling layers. The comparison on the hard subset (Table 3.8) shows that our system is better on SNLI, and reduces the difference on MultiNLI. We think that enriching their attention layer with n-grams such as ours, or, conversely, adding multi-head attention to our n-gram attention are promising directions for future research.

Regarding recursive encoders, the comparison to our method shows that using n-grams instead of syntax yields slightly better results (better on STS Benchmark by 5 points, worse on SICK-TS by 1 point) with less complexity. We think that these comparative results show that the n-gram attention model is able to partially capture syntactic information.

Finally, the system based on self-attention coupled with hard and soft attention has obtained slightly better results (same results on SNLI, 1 point better on SICK-TS). The use of self-attention is a promising direction of research, which could complement the good results of our n-gram attention model.

3.6 Conclusions and future work

In this work we extend attention models from pairs of words to pairs of word n-grams of variable length. We plugged our attention model on the well-known Decomposable Attention Model system (Parikh *et al.*, 2016), which is known for obtaining strong results on Natural Language Inference datasets. Our n-gram attention model improves results on five textual similarity and inference datasets, with up to 41% error reduction and 11 points of absolute

gain. The gains are especially large for datasets with small training data, and the hard subsets of MultiNLI and SNLI datasets (Gururangan *et al.*, 2018). Our experiments show that the alternative means to infuse context information into a word-to-word attention model (e.g. using a CNN or RNN over the context of occurrence) also improve results, but our method is the most effective. We also show that a trainable attention model increases results in all cases.

We think that the better results compared to recursive tree-based systems shows that n-grams are capturing some syntactic information. Our proposal can be seen as an intermediate step between learning a latent grammar Tai *et al.* (2015) and staying at the flat word level: we add some structure in the form of a flat set of possible word n-grams, but do not require a full-fledged tree.

From another perspective, our work can be additional evidence on the benefits of aligning chunks defended by Lopez-Gazpio *et al.* (2017). In their work, a linguistically motivated software identifies chunks and then aligns them across the target sentences. The system solving the task uses the provided training data to learn how to relate and align pairs of chunks. In our case, our n-gram attention model can be seen as inducing chunks (n-grams) and alignments between pairs of chunks without any direct supervision. We would like to explore whether chunk alignment corpora can be used to better train our n-gram attention model. Alternatively, our n-gram attention model might help improve systems solving the Interpretable STS task, including short answer grading (Riordan *et al.*, 2017).

Code and models are publicly available[13]. The analysis of state-of-the-art systems shows that our n-gram attention layer could be also beneficial (Chen *et al.*, 2016; Gong *et al.*, 2017; Shen *et al.*, 2018), as all top-scoring systems use word-to-word attention models. The benefits of our attention model could be also extended to other problems where the standard word attention model is used (Yang *et al.*, 2015; Luong *et al.*, 2015; Rajpurkar *et al.*, 2016). Finally, we would also like to explore whether the n-gram attention model trained in one task can be transferred to tasks with less training data.

[13] `https://github.com/lgazpio/DAM_Ngrams`

4. KAPITULUA

Esaldien arteko desberdintasunak topatzen eta azaltzen

Iñigo Lopez-Gazpio, Montse Maritxalar, Aitor Gonzalez-Agirre, German Rigau, Larraitz Uria and Eneko Agirre.
Interpretable semantic textual similarity: Finding and explaining differences between sentences.
Knowledge-Based Systems. 119, pp. 186 - 199. Elsevier.
ISSN: 0950-7051, Impact Factor: 3.325.
DOI: http://dx.doi.org/10.1016/j.knosys.2016.12.013, 2017.

Artikulu honetan antzekotasun semantiko interpretagarriaren deskribapen zehatza egiten da, geruza berri honen zehaztapenak azalduz. Artikulu honen helburu nagusia antzekotasun semantiko interpretagarriaren erabilgarritasuna bermatzea da, eta, horretarako, hainbat ebaluazio-eszenario definitzen dira hezkuntzaren alorra simulatuz. Inplementatutako sistemek lortutako emaitzek atazaren baliagarritasuna azpimarratzen dute irakaskuntzaren arlorako. Artikulu honek tesiaren 4. kapitulua osatzen du, eta antzekotasun semantiko interpretagarriaren inguruko lanen erreferentzia nagusia da.

Interpretable semantic textual similarity: Finding and explaining differences between sentences

I. Lopez-Gazpio, M. Maritxalar, A. Gonzalez-Agirre,
G. Rigau, L. Uria, E.agirre

IXA NLP group, Computer Science faculty, University of the Basque Country (UPV/EHU), Manuel Lardizabal 1, 20018, Donostia, Basque Country

Abstract

User acceptance of artificial intelligence agents might depend on their ability to explain their reasoning to the users. We focus on a specific text processing task, the Semantic Textual Similarity task (STS), where systems need to measure the degree of semantic equivalence between two sentences. We propose to add an interpretability layer (iSTS for short) formalized as the alignment between pairs of segments across the two sentences, where the relation between the segments is labeled with a relation type and a similarity score. This way, a system performing STS could use the interpretability layer to explain to users *why* it returned that specific score for the given sentence pair. We present a publicly available dataset of sentence pairs annotated following the formalization. We then develop an iSTS system trained on this dataset, which given a sentence pair *finds* what is similar and what is different, in the form of graded and typed segment alignments. When evaluated on the dataset, the system performs better than an informed baseline, showing that the dataset and task are well-defined and feasible. Most importantly, two user studies show how the iSTS system output can be used to automatically produce *explanations* in natural language. Users performed the two tasks better when having access to the explanations, providing preliminary evidence that our dataset and method to automatically produce explanations do help users understand the output of STS systems better.

4.1 Introduction

Since the early days of expert systems, it is acknowledged that one key factor for users and domain experts to accept expert systems in real-world domains is the ability of expert systems to explain their reasoning (Buchanan *et al.*, 1984; Lacave eta Dez, 2002; Korb eta Nicholson, 2010, p. 336). More recently, as machine learning systems are being deployed in society, interpretability of machine learning methods has become a hot topic of interest (Knight, 2016; Kim *et al.*, 2016a). We also think that user acceptance of artificial intelligence agents will depend on their ability to explain their reasoning. The challenge is to devise methods which allow the agents to explain *why* they take their decisions.

Our work explores interpretability in the context of Semantic Textual Similarity (STS) (Agirre *et al.*, 2012). STS measures semantic equivalence between two text snippets using graded similarity, capturing the notion that some pairs of sentences are more similar than other, ranging from no relation up to semantic equivalence. Systems attaining high correlations with gold, truth, scores have been routinely reported (Agirre *et al.*, 2012, 2014).

As an example of the STS task, when annotators judge the similarity between the following two sentences drawn from a corpus of News headlines, they define them as "roughly equivalent, but some minor information differs":

> 12 killed in bus accident in Pakistan
> 10 killed in road accident in NW Pakistan

Our final goal is to build Interpretable STS systems (iSTS systems for short) that are able to explain *why* they returned a specific similarity score, making explicit the differences and commonalities between the two sentences. The desired explanation for the two sample sentences would be something like the following:

> The two samples are strongly similar, because the two sentences talk about accidents with casualties in Pakistan, but they differ in the number of people killed (12 vs. 10) and level of detail: the first one specifies that it is a *bus* accident, and the second one specifies that the location is *NW* Pakistan.

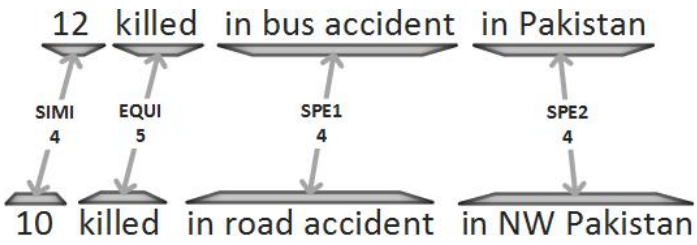


Figure 4.1: Representation of the interpretability layer. The two headlines were split in four segments each, which were aligned as follows: *"12"* is similar to *"10"* with a similarity score of 4 (SIMI 4), *"killed"* is equivalent to *"killed"* with score 5 (EQUI 5), *"in bus accident"* is more specific than *"in road accident"* with score 4 (SPE1 4), and *"in Pakistan"* is more general than *"in NW Pakistan"* with score 4 (SPE2 4). See Section 4.3 for more details on the annotation procedure.

While explanations come naturally to people, constructing algorithms and computational models that mimic human performance represents a difficult natural language understanding problem. In order to meet this challenge in the context of STS, we propose to add an interpretability layer on top of the STS system.

We build and evaluate an iSTS system that, given two sentences, returns a textual explanation of the commonalities and differences between the two sentences. The system formalizes the interpretability layer as an explicit alignment of segments in the two sentences, where alignments are annotated with a relation type and a similarity score. Figure 4.1 shows the formalization of the interpretability layer for the two sample sentences, including segments, alignments, types and scores of the alignments. Types include relations like equivalence, opposition, specialization, similarity or relatedness. The similarity scores for aligned segments range from 0 (no relation) to 5 (equivalence). The core part of the system is trained and evaluated on a dataset of sentence pairs which has been annotated with the alignments. Regarding the evaluation and feasibility of our proposal, the annotated dataset shows that the iSTS system performs better than an informed baseline, showing that the task is well-defined and feasible. The trained system is thus able to return the reasons for the similarity between the two sentences in the form of

typed segment alignments. This way, the final system will be able to explain to the user *why* the two sample sentences above were "roughly equivalent", producing text similar to the one shown above.

In addition to the dataset and core iSTS system, we also built a verbalization module, that is, a module that takes as input the alignments and produces a human-readable explanation based on templates. The system returns the following text for the alignment in Figure 4.1:

> The two sentences are very similar. Note that "in bus accident" is a bit more specific than "in road accident" in this context. Note also that "12" and "10" are very similar in this context. Note also that "in Pakistan" is a bit more general than "in NW Pakistan" in this context.

In order to measure the quality and usefulness of the explanations, direct comparison to human-elicited text (e.g. the explanation above) is problematic. Instead, we measure whether the automatically produced explanations are useful in two user studies. In the first study, English native speakers scored the similarity of sentence pairs, with and without automatically produced explanations. In the second study, we simulated a tutoring scenario where students were graded with respect to a reference sentence. The users, simulating to be students, had to state whether they agreed with the grade, with and without access to the automatically produced explanations. Both studies show that users that read the explanations agreed with the system scores more often than users which did not have access to explanations.

We summarize the contributions of this article as follows:

- It formalizes the interpretability layer in the context of STS as a graded and typed alignment between segments in the two sentences.
- It describes a publicly available dataset of sentence pairs (coming from news headlines and image captions) annotated with the interpretability layer following the above formalization.
- It describes a system that, given two sentence pairs, is able to return alignments between the segments in the two sentences annotated with relation type and a graded similarity score. The system is trained and

evaluated in the annotated dataset, with good results, well above an informed baseline and in the state-of-the-art.

- It presents an extension of the system which returns a textual explanation of the reasons for the similarity judgment.
- It shows two user-studies where the automatically produced explanations help users to better attain their tasks, providing preliminary evidence that our formalization and specific system are useful in real applications.

The dataset and core iSTS system have been previously reported in the Semeval 2015 workshop proceedings. The dataset was used in a subtask of the SemEval 2015 STS task (Agirre *et al.*, 2015a). The task description paper partially covered the dataset and participant systems. The system presented in this article is closely related to the system which participated in the task (Agirre *et al.*, 2015b). In this article we join the scattered parts and bring them together, adding motivation, a more detailed explanation of the annotation framework, statistics of the dataset, and the relation to other semantic annotation datasets. In addition, we present a novel verbalization module and two user studies which show the usefulness of interpretable STS.

Our research framework follows an empirical methodology. We first performed an analysis of related work (Section 4.2). We designed and annotated the iSTS dataset, a corpus of sentence pairs where annotators added an interpretability layer, as reported in Section 4.3, alongside an evaluation method, statistics, annotation quality indicators and a comparison to related datasets. We then designed and implemented an iSTS system which is able to annotate pairs of sentences with the interpretability information, alongside two baseline systems (Section 4.4). The system was developed on the train part of the iSTS dataset, and evaluated on the test part. The evaluation results, including error analysis and a comparison to the state-of-the-art is presented in Section 4.5. Having validated that the iSTS system is able to produce the interpretability layer with reasonable accuracy, we devised a method to map the interpretability layer into human-readable explanations (the verbalization module), and evaluated those explanations on two user-studies (Section 4.6).

4.2 Related work

Early work on adding explanations in the context of bayesian networks includes both visualizations and verbal explanations about the model itself or the conclusions drawn about the domain (Cooper, 1984; Suermondt, 1992). For instance, Elvira (Consortium, 2002) is a Bayesian Network package that offers both verbal explanations (about nodes and arcs) as well as graphical explanations.

Explanations are important in the teaching domain, where Intelligent Tutoring Systems (ITS) strive to provide feedback beyond correct/incorrect judgments. In most cases, the systems rely on costly domain-dependent and question-dependent knowledge (Aleven *et al.*, 2001; Jordan *et al.*, 2006), but some scalable alternatives based on generic Natural Language Processing (NLP) techniques are also available (Nielsen *et al.*, 2009). Our approach is related in spirit with this last paper, but we formalize the interpretability layer differently, as we will see below.

In the area of NLP, the interpretability of representation models learned from raw data is also a widespread concern. Ritter *et al.* (2010) show that they are able to infer classes which are easily interpretable by humans, and Fyshe *et al.* (2015) argue that the dimensions of their word representations correspond to easily interpretable concepts. To our knowledge this article is the first research work in the area of NLP addressing explicit, human-readable, explanations about the semantic similarity of two sentences.

Our work is situated in the area of Natural Language Understanding, where two related enabling tasks have been extensively used to evaluate the quality of semantic representations, STS and textual entailment. STS has been the focus of several SemEval tasks starting in 2012 (Agirre *et al.*, 2012) and ongoing at the time of writing this paper[1]. Given a pair of sentences, s1 and s2, STS systems compute how similar s1 and s2 are and return a similarity score. STS is related to both paraphrasing and textual entailment, but instead of being binary it reflects a graded notion. It also differs from textual entailment in that it is not directional (Giampiccolo *et al.*, 2007;

[1] `http://ixa2.si.ehu.eus/stswiki`

Bowman *et al.*, 2015). STS is an enabling technology with applications in Machine Translation evaluation, Information Extraction, Question Answering and Text Summarization. Our work reuses existing STS datasets, and adds an interpretable layer, in the form of typed alignments between sentence segments.

Systems performing STS have reported the use of several NLP tools (Agirre *et al.*, 2015a), such as lemmatizers, part of speech taggers, syntactic parsers, name entity recognizers, and distributional or knowledge-based resources like ConceptNet (Liu eta Singh, 2004), PPDB (Ganitkevitch *et al.*, 2013), WordNet (Fellbaum, 1998) and word embeddings (Mikolov *et al.*, 2013a). As an example, top performing runs at the SemEval 2015 STS task were based on supervised systems taking into account word alignment ratios as well as compositional sentence vectors to compute the final similarity score (Sultan *et al.*, 2015; Hänig *et al.*, 2015).

More recently, recurrent neural networks have obtained promising results. Tai *et al.* (2015) use Long Short-Term Memory networks (Hochreiter eta Schmidhuber, 1997b), which are able to preserve sequence information effectively during input processing, and exploit the syntactic properties that are inherent to natural language at the same time. Alternatively, He eta Lin (2016) use convolutional neural networks to explicitly model pairwise word interactions, proposing a similarity focus mechanism to identify important correspondences for better similarity measurement.

Our formalization of the interpretable layer proposes the explicit alignment of segments, where each alignment is labeled with a relation type and a similarity score. Previous work on semantic alignment between text segments in the same language[2] have usually focused on the word level, with some exceptions. For instance, Brockett (2007) released the 2006 PASCAL corpora composed of sentence pairs, where semantically equivalent words and phrases in the Text (T) and Hypothesis (H) sentences were aligned. Each word of H was either linked to one or more words of T or it was left unlinked, and the links were marked as either sure or possible depending on the degree of confidence in the alignment. Annotators of the dataset viewed the sentence

[2]As opposed to alignment of parallel corpora in machine translation settings.

pairs of the corpora as pairs of parallel strings of words with lines of association between them, with limited coverage of some phrases like multiword expressions. In our work we go one step further and focus on text segments beyond words, as well as adding alignment types and similarity scores. In a similar effort, Rus *et al.* (2012) aligned tokens from a STS dataset, including some short phrases, such as chunks which were semantically equivalent but non-compositional. In our case our formalization covers all kind of segments, including non-equivalent and equivalent segments, compositional or not.

From another perspective, our dataset is related to work on semantic compositionality, that is, how the semantic representation of syntactic components is built based on the semantic representation of the words. As we just mentioned, word alignment datasets do exist (Brockett, 2007; Rus *et al.*, 2012), but they are not amenable to studies in semantic compositionality. Current similarity and entailment datasets either focus on the word level, e.g. word similarity and relatedness datasets (Finkelstein *et al.*, 2002; Hill *et al.*, 2015), or on the sentence level, e.g. STS datasets (Agirre *et al.*, 2012) and entailment datasets (Giampiccolo *et al.*, 2007; Bowman *et al.*, 2015). Our dataset fills an important gap, as it provides a testing ground for semantic representation and compositionality methods at the chunk level.

In recent work, which has been performed in parallel to ours, Pavlick *et al.* (2015) annotated an automatically derived database of paraphrases for short phrases (Ganitkevitch *et al.*, 2013) with entailment relations from Natural Logic (MacCartney eta Manning, 2008). They used crowdsourcing to annotate by hand around 14 thousand phrase pairs in the database. Section 4.3 includes a head-to-head comparison of the annotation schemes, showing that our work is closely related and complementary.

In a different strand of work coming from the educational domain and close to textual entailment, Nielsen *et al.* (2009) defined so-called facets, where each facet was a pair of words and a non-explicit semantic relation between both words. Each facet in the hypothesis text, usually a sentence, is annotated with information of whether it is entailed by the reference text. In the context of tutoring systems, their dataset comprises student responses and reference answers. Each reference answer was decomposed by hand into

its constituent facets. The student answers are annotated with a label for entailed facets of the corresponding reference answer, but, contrary to our proposal, there is no explicit alignment between facets, and the facets do not necessarily correspond with text segments, but rather represent pairs of words having an unknown semantic relation in the text. Our initial motivation for interpretable STS was similar to that of Nielsen *et al.* (2009), as we think interpretability is especially useful in the field of tutoring systems, but we depart from that work in explicitly aligning segments in both sentences, as well as providing labels for the relation and similarity scores.

The idea of facets was later followed by Levy *et al.* (2013), which call it partial textual entailment. This approach is complementary to ours, in that they could also try to align facets and characterize the semantic relations as well as the alignment relations. From another perspective, the same way they enrich textual entailment datasets with partial entailment annotations, we also enrich STS datasets with explicit alignments, where our types are related to entailment relations.

Other related work includes a SemEval task related to tutoring systems that automatically score student answers, the Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge (Dzikovska *et al.*, 2013). This task was the first large-scale and non-commercial automatic short answer grading competition (Burrows *et al.*, 2015). The goal of the mentioned task was to assess student responses to questions in the science domain, focusing on the correctness and completeness of the response content. In a typical scenario, they expected that a correct student answer would entail the reference answer. The goal of the mentioned task was to label the student answers according to different categories (i.e. correct, partially correct or incomplete, contradictory, irrelevant and out-of-domain). The task included a pilot subtask where participants had to annotate facets. In our opinion, effective feedback needs to identify the specific text segments of the student answers that differ from the reference answer, which we do via alignments.

The task of finding semantic correspondences between pairs of texts is also related to plagiarism detection, where the system needs to detect passages of text that have been copied verbatim or paraphrased from other docu-

ments (Lukashenko *et al.*, 2007). Plagiarized passages usually go beyond the sentence level, encompassing several sentences or paragraphs. Recent datasets comprise pairs of passages where one is the original and the other one is the plagiarized passage (Potthast *et al.*, 2015). These pairs have been created either automatically or manually, asking volunteers to manipulate the original passage. The focus of our work is on naturally occurring pairs of sentences, but, in the future, plagiarism datasets could be tapped as an additional source of sentence pairs to be annotated using our framework. Regarding plagiarism detection techniques, the main difference is that they need to scale up in order to explore the similarity between all passages of the suspicious document and all passages in the document base, adding a severe computational constraint on the techniques that can be used, and leading to shallow techniques closely related to information retrieval (Potthast *et al.*, 2010a; Barrn-Cedeo *et al.*, 2013).

4.3 Building the iSTS dataset

This section presents the iSTS dataset. We first introduce the annotation procedure, followed by the source of the sentence pairs, the evaluation method, and inter-tagger annotation data.

Annotation procedure

We briefly introduce the annotation procedure, which is fully documented in the annotation guidelines [3]. Given a pair of sentences, the procedure to be followed by annotators is the following:

1. First of all, the annotator identifies the chunks in each sentence separately, regardless of the corresponding sentence in the pair.
2. Secondly, the annotator aligns the chunks in order, from the clearest and strongest correspondences to the most unclear or weakest ones.
3. Third, for each alignment, the annotator provides a similarity score.
4. Finally, the annotator chooses the label or tag for each alignment.

[3] `http://alt.qcri.org/semeval2015/task2/data/uploads/annotation_guidelines_semeval-2015_task2_interpretablests.pdf`

```
Noun phrases: [The girl] / [Bradley Cooper and JJ Abrams]
Verb chains: [is arriving] / [does not like]
Prepositional phrases: [at a time] / [the house] [of that man]
Adverbial phrases: [of course]
Other expressions: [once upon a time] / [by the way]
```

Figure 4.2: Examples of chunks.

Text segments. Segments are annotated according to the definition of *chunks* (Abney, 1991): "a non-recursive core of an intra-clausal constituent, extending from its beginning to its head. A typical chunk consists of a content word surrounded by a constellation of function words, matching a fixed template". When marking the chunks of each sentence, the annotator follows the CONLL 2000 task guidelines[4], which were adapted slightly for our purpose: The main clause is split in smaller chunks consisting on noun phrases, verb chains, prepositional phrases, adverbs and other expressions. Figure 4.2 shows some examples of chunks. In order to help the annotators, we previously run the sentences through a publicly available open-source chunker[5] trained on CONLL 2000 corpora (Agerri *et al.*, 2014).

Alignment. The alignment is performed using devoted interface [6]. When aligning, the meaning of the chunks in context are taken into account. Annotators must try to align as many chunks as possible. Given some limitations in the interface, we decided to focus on one-to-one alignments, that is, one chunk can be aligned with at most one chunk. For this reason, when having two options to align, only the strongest corresponding chunk will be aligned. The other chunk(s) will be left *unaligned*, and labeled with *ALIC*. Chunks can be also left unaligned if no corresponding chunk is found (*NOALI* label). Punctuation marks are ignored, and left unaligned.

Score. Once chunks are aligned, the annotator provides a similarity score for each alignment, where the score ranges from 5 (maximum similarity,

[4] `http://www.clips.ua.ac.be/conll2000/chunking/`

[5] `https://github.com/ixa-ehu/ixa-pipe-chunk`

[6] We modified a tool developed by LDC to align words `https://www.ldc.upenn.edu/language-resources/tools/ldc-word-aligner`. We reused their XML-based annotation format as well.

Label	Chunk1	Chunk2	Score
EQUI	abduct	kidnapped	5
OPPO	soar	slump	4
SPE1	two mountain goats	two animals	1
SPE2	in Pakistan	in NW Pakistan	4
SIMI	Russia	South Korea	3
REL	on the porch	on a couch	2

Table 4.1: Examples of aligned chunks, with label and score.

equivalence) to 0 (no relation at all). Note that an aligned pair would never score 0, as that would mean that the two chunks should not be aligned. See below for further restrictions concerning possible score values for specific labels.

Label. When assigning labels to aligned chunks, the interpretation of the whole sentence, including common sense inference, has to be taken into account. The possible labels are the following:

EQUI, both chunks are semantically equivalent;

OPPO, the meanings of the chunks are in opposition to each other;

SPE1 (or SPE2), chunk in sentence 1 is more specific than chunk in sentence 2 (or vice versa);

SIMI, the meaning of the chunks are similar, and the chunks are not *EQUI*, *OPPO*, *SPE1*, or *SPE2*;

REL, the meaning of the chunks are related, but they are not *SIMI*, *EQUI*, *OPPO*, *SPE1*, or *SPE2*;

These six labels are mutually exclusive, and each alignment should have one and only one such label. In addition, the following optional labels can be used in any alignment:

FACT, the factuality, i.e. whether the statement is or is not a fact or a speculation is different in the aligned chunks.

POL, the polarity, i.e. the expressed opinion (positive, negative or neutral) is different in the aligned chunks.

Note that *ALIC* and *NOALI* can also be *FACT* or *POL*, meaning that the respective chunk adds a factuality or polarity nuance to the sentence. After annotating scores and labels, the annotator should see that the following constraints are enforced:

- *NOALI* and *ALIC* should not have scores.
- *EQUI* should have a 5 score.
- The rest of the labels should have a score larger than 0 and lower than 5.

iSTS	NL
EQUI	≡
OPPO	¬
SPE1	⊏
SPE2	⊐
SIMI	~
REL	~

Table 4.2: Relation between our alignment types and the Natural Logic entailment relations used by Pavlick *et al.* (2015). All relations are one-to-one, except our SIMI and REL, which are both conflated as ~ in natural logic

Discussion. Some examples of labels and scores are provided in Table 4.1. For EQUI we see that the score needs to be 5. In the case of OPPO, the two pairs are assigned a high score. We instructed the annotators to assign high scores to strong opposites, following the approach in the most commonly-used evaluation gold standard for semantic models for word, WS-353 (Finkelstein *et al.*, 2002). In their instructions they stated the following: "... when estimating similarity of antonyms, consider them similar (i.e., belonging to the same domain or representing features of the same concept), not dissimilar". Their approach had the shortcoming of not being able to differentiate synonyms from antonyms. Our annotation solves this issue, as we assign antonyms a high similarity score and the OPPO label.

Regarding the difference between similar (SIMI) and related (REL) chunks, this distinction stems from psychology-related studies on the relation between

words (Hill *et al.*, 2015). Similarity refers to conceptually similar concepts or entities which share many features and can be categorized in the same category (Tversky, 1977). For instance, Russia and South Korea share many features, and can be categorized as countries. In another example, couch and sofa are also similar, as they are both pieces of furniture with many features in common. Relatedness refers to concepts which are closely related but do not necessarily share features (Hill *et al.*, 2015), where they call it *association*. For instance Putin is the president of Russia, and thus both entities are related. In another example, porch and couch are related, as a couch is sometimes located in a porch. Note that, in general, relatedness also encompasses similarity, but we only assign the REL label to those pairs of chunks which are not similar, making both REL and SIMI exclusive labels.

Our labels are closely related to those used in Natural Logic (MacCartney eta Manning, 2008), and later adapted for the purpose of annotating a database of paraphrases (Pavlick *et al.*, 2015). We compare our annotations to the latter, as it is closer to this work. They use a set of six mutually exclusive entailment relations:

- Equivalence (couch $\equiv$ sofa)
- Opposites (old $\neg$ young)
- Forward / backward entailment (crow $\sqsubset$ bird)
- Related by something other than entailment (boy $\sim$ little)
- Unrelated (professor # cucumber)

Our labels have been created independently from theirs, but the overlap between both annotation schemes is remarkable. Table 4.2 shows the mapping between the respective labels, which is a one-to-one mapping with exception of their $\sim$ relation, where we distinguish between similarity and relatedness. Several researchers have argued about the convenience to separate these phenomena (Agirre *et al.*, 2009; Hill *et al.*, 2015), and we thus think that our labels add a valuable piece of information.

Our annotated resource is thus related and complementary to that released by Pavlick *et al.* (2015), who annotated with their relation label (including no-relation) a subset of an automatically derived paraphrase database

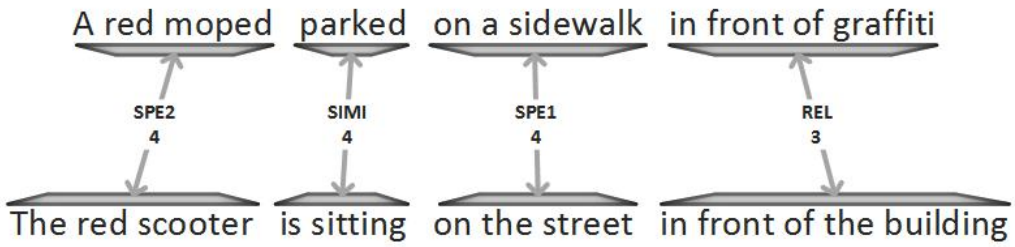


Figure 4.3: Two sentences from the Images dataset. The two sentences from the Images dataset were split in four segments each, which were aligned as follows: *"A red moped"* is more general than *"The red scooter"* with a score of 4 (SPE2 4), *"parked"* is similar to *"is sitting"* with score 4 (SIMI 4), *"on a sidewalk"* is more specific than *"on the street"* with score 4 (SPE1 4), and *"in front of graffiti"* is related to *"in front of the building"* with score 3 (REL 3). See Section 4.3 for more details on the annotation procedure.

(Ganitkevitch *et al.*, 2013). In our case, we annotate both relation label and score of manually identified chunks in text pairs. In addition, the annotator has checked that the unaligned chunks are not related to any of the chunks in the corresponding sentence, and we thus implicitly annotated pairs of chunks having no relation (NOALI label). Note that the source of pairs in each resource is different: while they labeled pairs of phrases which had been automatically induced as being paraphrastic, we label pairs of chunks in naturally occurring sentences from different similarity ranges. In addition, we distinguish between similar and related pairs, and label explicitly factuality and polarity phenomena.

Source of the dataset

The dataset comprises pairs of sentences from news headlines (Headlines) and image descriptions (Images), as gathered for the STS competition (Agirre *et al.*, 2012, 2014). We already showed a sample pair from Headlines (cf. Figure 4.1), and Figure 4.3 shows a sample pair from Images, together with their alignment. The Headlines corpus is composed of naturally occurring

news headlines gathered by the Europe Media Monitor engine from several different news sources (from April 2nd, 2013 to July 28th, 2014) as described by Best *et al.* (2005). The Images dataset is a subset of the PASCALVOC-2008 dataset, as described by Rashtchian *et al.* (2010), which consists of 1000 images with around 10 descriptions each. The dataset comprised 756 and 750 sentence pairs from Headlines and Images, respectively. We reused the sentence pairs released by the STS Semeval task, which include both similar and dissimilar sentence pairs following a uniform distribution [7]. The dataset is split evenly in training and testing subsets, and is freely available [8].

Table 4.3 describes the statistics for the Headlines and Images datasets. Headlines contain slightly less chunks and less tokens per chunk than image captions. More than half of the aligned pairs in both datasets have a score of 5 (which corresponds to EQUI pairs) with a decreasing number of aligned pairs for each score range. Regarding the labels, EQUI is the most used label, followed by SIMI, SPE1 and SPE2, REL and OPPO. The breakdown in scores and types is very similar in both datasets. ALIC is used a few times, more often in the Headlines dataset. There is a large number of unaligned chunks, which is natural, given that some pairs of sentences have medium and low similarity. Up to 58% of the chunks are aligned in the Images dataset and 72% in headlines. Finally, FACT and POL are seldom used in the news dataset, and never in the Images dataset.

The dataset contains a wealth of information for the 1506 sentence pairs, including chunking, 8437 aligned chunks with a score and relation label, and 20400 pairs of chunks which have no relation [9]. The amount of information in size and richness is comparable, and in some cases larger, than those made available in related work (cf. Section 4.2). For instance, the STS evaluation in 2012 (Agirre *et al.*, 2012) included 5250 pairs with a score. The alignments in

[7] Please see (Agirre *et al.*, 2014) for details on the sampling procedure.

[8] `http://alt.qcri.org/semeval2015/task2/data/uploads/sts2015-interpretability-train.v3.tgz` and `http://alt.qcri.org/semeval2015/task2/data/uploads/test_evaluation_task2c.tgz`

[9] This number is derived from pairing the unaligned chunks with all chunks in the corresponding sentence, as the annotator has checked that there is no relation between them (NOALI label).

(Brockett, 2007) comprise 24601 alignment decisions, and Rus *et al.* (2012) aligned with no labels tokens in 700 sentence pairs. Pavlick *et al.* (2015) annotate 14000 pairs of chunks with a label, but include many pairs with no relation. This is lower than our 28837 aligned and unaligned pairs, and, for those that are aligned we provide both label and score.

The low number of OPPO, FACT and POL labels stems from several factors. First of all, from the fact that we reused the STS datasets, which include pairs of sentences without biasing them towards particular linguistic phenomena. Another factor is that the datasets are based on captions and news headlines, so the low number follows the natural distribution of those phenomena in those corpora. This is a shortcoming of the method to collect pairs of sentences used by the STS authors. The solution is not clear. For instance, other datasets like (Bentivogli *et al.*, 2016) have manually manipulated existing sentences to produce variants which exhibited linguistic phenomena of interest. The bias introduced by the manipulations could be replicated by the systems using ad-hoc heuristics, reducing the interest of the dataset. Thus, gathering non-artificial pairs of sentences (i.e. sentence pairs free of manual manipulation) that exhibit phenomena like OPPO, FACT and POL is an interesting research problem in its own.

Annotation Effort

The annotation of 1501 pairs took 70 hours, around 2.8 minutes per sentence pair. The annotation was faster towards the end of the project, at around 2.3 minutes per pair. The annotation interface was key to allow for fast annotation.

Evaluation measures

In order to evaluate iSTS systems, we adopt word alignment evaluation methods from the Machine Translation community. In particular, the evaluation method is based on that of Melamed (1998), which uses the F1 of precision and recall of token alignments. Note that Fraser eta Marcu (2007) argued that F1 is a better measure than Alignment Error Rate. The idea is that segment alignment is mapped into token alignment, where all token pairs in

	Headlines				Images			
	Train	Test	All	%	Train	Test	All	%
Sentence pairs	378	378	756		375	375	750	
Chunks/sentence	4.2	4.2	4.2		4.5	4.5	4.5	
Tokens/chunk	1.9	1.9	1.9		2.2	2.3	2.25	
Aligned pairs	1064	1102	2166		969	942	1911	
Score $\in$ [5]	652	665	1317	60.8	529	499	1028	53.8
Score $\in$ [4,5)	189	225	414	19.1	247	268	515	26.9
Score $\in$ [3,4)	133	126	259	12.0	101	107	208	10.9
Score $\in$ [2,3)	80	70	150	6.9	75	55	130	6.8
Score $\in$ [1,2)	10	16	26	1.2	17	13	30	1.6
EQUI	652	665	1317	60.8	529	499	1028	53.8
SPE1	98	99	197	9.1	108	126	234	12.3
SPE2	86	108	194	8.9	129	109	238	12.4
SIMI	171	154	325	15.0	174	170	344	18.0
REL	48	66	114	5.3	29	35	64	3.3
OPPO	9	10	19	0.9	0	3	3	0.2
ALIC	92	99	191		53	39	92	
NOALI	949	841	1790		1406	1468	2874	
FACT	10	20	30		0	0	0	
POL	3	0	3		0	0	0	

Table 4.3: Headlines and Images dataset statistics across splits. The first three rows report, respectively, the number of sentence pairs, chunks per sentence and tokens per chunk. The rows below report the number of aligned chunk pairs, with a break-down according to the similarity score, followed by a breakdown according to the label of aligned pairs. The last four rows report the number of unaligned chunks (ALIC, NOALI), and how many times the additional FACT and POL labels are used.

the aligned pairs are aligned with some weight. The weight of each token-token alignment is the inverse of the number of alignments of each token, the so-called fan out factor (Melamed, 1998). Precision is measured as the ratio of token-token alignments that exist in both system and gold standard files, divided by the number of alignments in the system. Recall is measured similarly, as the ratio of token-token alignments that exist in both system and

	Headlines	Images
Alignment	91.4	100
Type	77.1	81.0
Score	84.8	95.2
Type + Score	73.8	77.1

Table 4.4: Inter-tagger agreement (%) on the Headlines and Images dataset, using the evaluation method in Section 4.3.

gold-standard, divided by the number of alignments in the gold standard. Precision and recall are evaluated for the alignments of all pairs in one go.

The evaluation is done at four different levels: segment alignment alone (ignoring labels and scores), segment alignment where we require that labels agree (i.e. pairs of segments with different labels are ignored), segment alignment where differences in score are penalized, and, finally, segment alignment score where both labels and scores are taken into account. The later is the overall evaluation criteria (i.e. Type+Score), as it takes into account the full task. The evaluation script is freely available together with the dataset.

Quality of annotation

To measure the viability and quality of the annotation and to calculate the *inter-tagger agreement* (ITA), two annotators annotated, individually, a random subset of 20 sentence pairs, 10 from each dataset. The 20 sentences contain 363 tokens, 180 chunks and 130 alignments. The evaluation dataset thus comprises 363 chunking decisions, 180 alignment decisions, and 260 labeling and scoring decisions (two for each aligned pair), totaling 803 items of evaluation. Before starting the annotation, both annotators read the guidelines and discussed any issue they could find. The agreement was computed using the evaluation script (T+S, cf. Section 4.3), where one tagger was taken as the system and the other one as the gold standard. This method to evaluate the agreement allows for head-to-head comparison to the performance of systems, where the ITA should set the upperbound for systems. Overall results for the agreement are shown in table 4.4.

The segment alignment is done with very high agreement (over 90%), both

for Headlines and Images. The agreement on type is also high, around the 80%, as well as the agreement on scores (over 80%). When considering the agreement on both type and score, the scores are also over 70%, with the highest score for the simpler Images dataset. The high results show that the annotation task is well-defined and replicable.

4.4 Constructing an iSTS system

A system for Interpretable STS needs to perform chunking, align the chunks, label and score the alignments. We first describe a baseline system which performs each of the steps in turn, and then present some improvements.

Baseline system

The baseline performs each of the steps using some publicly available algorithms. It first runs the ixa-pipes chunker[10] (Agerri *et al.*, 2014). We then lower-case all tokens and align identical tokens. Chunks are aligned based on the number of aligned tokens in a greedy manner, starting with the pair of chunks with the highest relative[11] number of aligned tokens. Chunks with no aligned tokens are left unaligned. Finally, the baseline uses a rule-based algorithm to directly assign labels and scores, as follows: aligned chunk pairs are assigned the *EQUI* label (the majority label in the dataset), and the rest are either assigned *ALIC* (if they contain aligned tokens), or *NOALI* (if they do not contain aligned tokens). The procedure to assign scores follows the alignment guidelines: *EQUI* pairs are scored with the maximum score and the rest are scored with *0*.

Chunking

Given that the chunker is not perfect, we analyzed the output of the chunker with respect to the gold chunks available in the training data, and used some regular expressions to improve chunking. The rules concern conjunctions,

[10] `https://github.com/ixa-ehu/ixa-pipe-chunk`

[11] The mean number of tokens is used for normalization.

punctuation symbols and prepositions, where the rules are used to join adjacent chunks. The rules mainly join prepositions and noun phrases into a single chunk, as well as noun phrases separated by punctuation or conjunctions, or a combination of those. In addition to the chunker, we run the Stanford NLP parser (Klein eta Manning, 2003), producing part of speech, lemma and dependency analysis.

Alignment

We use a freely available state-of-the-art monolingual word aligner (Sultan *et al.*, 2014) for producing token alignments. In order to produce the chunk alignment, each possible chunk alignment is weighted according to the number of aligned tokens in the chunks. The Hungarian-Munkres algorithm (Munkres, 1957) is then used to find the chunk alignments which optimize the overall alignment weight.

Labeling

Alignments are labeled using a multiclass supervised classification algorithm, trained with positive alignments in the training data [12]. We use twenty one features including token overlap, chunk length, WordNet similarity between chunk heads and WordNet depth. The features are listed in table 4.5.

We used Support Vector Machines (Chang eta Lin, 2011). As training data is limited, we performed grid search to optimize the cost and gamma parameters using randomly shuffled 5-fold cross validation. In these development experiments we found that the classifier was failing to detect FACT and POL, so we removed these labels from the training in the final system.

The development experiments also showed that the performance of the classifier was sensitive to the quality of the chunker. The classifier was first trained and tested using cross-validation on data which contained gold chunks and gold alignments, but when we run the classifier on test folds which contained system chunking, the performance suffered. We tried a data-augmentation method, where we used automatically produced chunks combined with the

[12] We extracted all aligned pairs with *EQUI*, *OPPO*, *SPE1*, *SPE2*, *SIMI* and *REL* labels.

#	Feature description
1	Jaccard overlap of content words
2	Jaccard overlap of non stopwords
3	Jaccard overlap of stopwords
4	Difference in length betweeen chunks 1 and 2
5	Difference in length betweeen chunks 2 and 1
6	Max WordNet path similarity of sense pairs (Pedersen *et al.*, 2004)
7	Max WordNet LCH similarity of sense pairs (Leacock eta Chodorow, 1998)
8	Max WordNet JCN similarity of sense pairs (Jiang eta Conrath, 1997)
9	Same as 6 but simulating root with the maximum common subsumer
10	Same as 7 but simulating root with the maximum common subsumer
11	Same as 8 but simulating root with the maximum common subsumer
12	Whether chunk 1 senses are more specific than chunk 2 senses in the WordNet hierarchy (Fellbaum, 1998)
13	Whether chunk 2 senses are more specific than chunk 1 senses in the WordNet hierarchy
14	Difference in WordNet depth of segment head
15	Minimum value of pairwise difference of WordNet depth
16	Maximum value of pairwise difference of WordNet depth
17	Lemmatized lowercased tokens of chunk 1
18	Lemmatized lowercased tokens of chunk 2
19	Maximum similarity value using first resource in Section 4.4
20	Maximum similarity value using second resource in Section 4.4
21	Maximum similarity value using third resource in Section 4.4

Table 4.5: Features used by the supervised classifier to assign labels to aligned chunk pairs.

gold standard chunks to train the labeling system. We tried several variations via cross-validation on the train dataset, obtaining the best results using a concatenation of the gold dataset (with gold chunks) and a version of the gold dataset which mixed the automatically produced chunking with the gold alignments and labels. We thus trained the final classifiers on this data-augmented dataset. For instance, we show below a training sentence (cf. Figure 4.3) with the gold standard chunks (first sentence). We augment the training with the same example tagged with the automatically produced

chunks (second sentence), with errors like chunking *moped* in a verb group and *front* in a noun phrase. Below we omit the alignment and labeling, for easier presentation.

[A red moped] [parked] [on a sidewalk] [in front of graffiti]
*[A red] [moped parked] [on a sidewalk] [in front] [of graffiti]

Scoring module

The scoring module uses a variety of word similarity resources, as follows:

- Euclidean distance between Collobert and Weston Word Vector (Collobert eta Weston, 2008). The distances d were converted to similarity s in the [0..1] range using the following formula, $1 - d/max(D)$, where D contains all distances observed in the dataset.
- Euclidean distance between Mikolov Word Vectors (Mikolov *et al.*, 2013a). The distance was converted into similarity as above.
- PPDB Paraphrase database values (Ganitkevitch *et al.*, 2013). We used the XXXL version. This resource yields conditional probabilities. As our scores are undirected, when the database contains values for both directions, we average.

Given a pair of aligned chunks (C_1 and C_2), we compute the similarity for any word pair $sim(w, v)$ in the chunks, where $w \in C_1$ and $v \in C_2$, as the maximum of the similarities according to the three resources above. We then compute the similarity between the chunks as the mean of two similarities, the addition of similarities for each word in the first chunk and the addition of similarities for each word in the second chunk, as follows:

$$sim(C_1, C_2) = \frac{1}{2}\left(\frac{\sum_{w \in C_1}(max_{v \in C_2} sim(w, v) * idf(w))}{\sum_{w \in C_1} idf(w)}\right. \quad (4.1)$$

$$\left. + \frac{\sum_{w \in C_2}(max_{v \in C_1} sim(w, v) * idf(w))}{\sum_{w \in C_2} idf(w)}\right) \quad (4.2)$$

	Headlines				Images			
	ALI	TYPE	SCORE	T+S	ALI	TYPE	SCORE	T+S
BASE	67.0	45.7	60.7	45.7	70.6	37.0	60.9	36.9
BASE+	**77.1**	50.2	68.9	50.2	**83.9**	44.5	72.8	44.5
FULL	**77.1**	**53.4**	**70.1**	**52.2**	**83.9**	**60.9**	**76.1**	**58.8**
$\text{FULL}_{GChunks}$	89.9	64.0	82.1	61.9	88.5	65.6	80.9	61.6

Table 4.6: Results (F1 %) of our three systems on each of the datasets. Columns show the results on each evaluation criteria, where ALI stands for alignment, TYPE for the label, SCORE for the scoring and T+S for both type and score. Best results in bold. The last row shows the results for the FULL system when using gold standard chunks instead of automatically produced chunks.

In the equation above *idf* is the inverse document frequency, as estimated using Wikipedia as a corpus.

Three systems

We developed three systems: the baseline (BASE, cf. Section 4.4), an improved baseline (BASE+) with improved chunking and alignment models (cf. Sections 4.4 and 4.4) but the same labeling and scoring modules as the baseline, and the full system (FULL) with supervised labeling and similarity-based scoring (Sections 4.4 and 4.4). The systems were developed using the training subset of the dataset alone, with no access to the test.

4.5 Evaluation

This section explains the results of the three systems we have developed and an error analysis of them. Then, a comparison with respect to the state-of-the-art is presented.

Developed systems

We evaluated the three systems (BASE, BASE+ and FULL) according to the evaluation measures set in Section 4.3. Table 4.6 shows the results on the Headlines and Images datasets. The better chunking and alignment (BASE+

and FULL) improves the alignment F1 score more than 10 points in both datasets with respect to BASE. The poor performance in alignment causes the baseline system to also attain low F1 scores in type and score, as well as the overall F1 score (T+S). The comparison between BASE+ and FULL shows that the classifier is able to better assign types, specially for Images. The method to produce the score is also stronger in FULL, and thus produces the best overall F1 (T+S). Note that the performance for the four available metrics decreases, as all metrics are bounded by ALI, and T+S is bounded by both TYPE and SCORE.

All in all, the alignment results are strong, but the decrease of performance when taking into account the type shows that this is the most difficult task right now, with score being an easier task. In fact, had the labeling been perfect, the TYPE F1 score would be the same as ALI F1 score, but a drop around 23 absolute points is observed in both datasets, while scoring performance only drops around 7 absolute points with respect to ALI. Regarding the two datasets, Headlines are more challenging, with lower scores across the four evaluations.

Error analysis

We performed an analysis of the errors performed by the FULL system at each level of processing, starting with chunking. The last row in Table 4.6 reports the results of the FULL system when running on gold standard chunks. The results improve for both datasets, with very high alignment results, and show that chunking quality is key for good performance. The results when using gold chunks are comparable for the two datasets, which indicates that the difference in performance for Headlines and Images when running the system on raw data (first three rows in Table 4.6) is caused by the automatic chunker. We can thus conclude that Headlines is more difficult to chunk than Images, which causes worse performance on this dataset. Some of the errors in chunking seem to be related to verbs, as shown in an example below, and could be caused by the particular syntactic structures used in news headlines, which are different from those expected by the automatic chunker.

* [Three] [dead] [**after helicopter crashes**] [into pub]

[Three] [dead] [**after helicopter**] [**crashes**] [into pub]

Regarding the quality of the alignments, we found that the aligner tended to miss some alignments because it did not have access to semantic relations between words (e.g. cows and horse below) or numbers (500 and 580 below). The following pairs include, in bold, chunks which should have been aligned by the system:

Two cows graze in a field.
A brown horse in a green field.

Bangladesh building disaster death toll passes **500**
Bangladesh building collapse: death toll climbs to **580**

In order to check type-labeling errors, we built a confusion matrix between the FULL system and the gold standard for the Headlines dataset (see left of Figure 4.4). The confusion matrix was built with correct alignments, as incorrectly aligned tokens cannot be analyzed for type errors.

Most errors of the system are caused by the system being biased to return EQUI, which we think is caused by the imbalance of the classes in train (cf. Table 4.3). For example, in the next pair of sentences the aligned chunks **(in bold)** should have been labeled as SPE1 instead of EQUI.

Asiana jet crash lands at San Francisco airport
Plane crash lands at San Francisco airport

In some cases, the system is not able to label equivalent chunks due to mistakes when recognizing identical entities or synonyms. In the next examples, the chunks in sentence 2 shown **in bold** have been labeled as more specific than the corresponding chunks of sentence 1 (also **in bold**). However, in both cases the alignments should have been labeled as EQUI instead of SPE2.

Matt Smith **to leave** Doctor Who after 4 years
Matt Smith **quits** BBC's Doctor Who

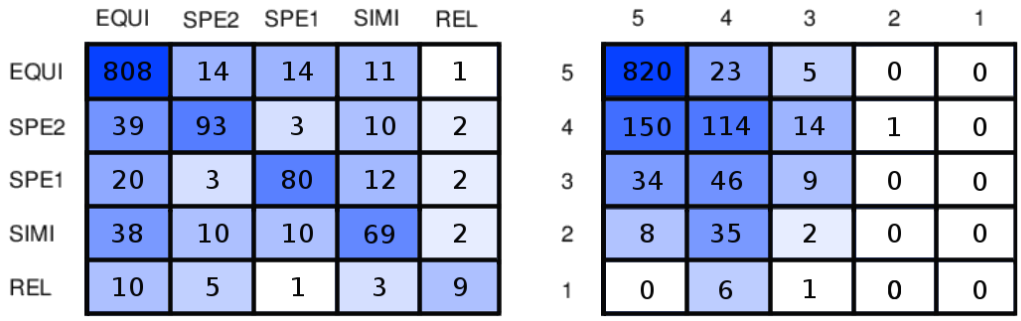


Figure 4.4: Label and score confusion matrices as heatmaps between the FULL system and the gold standard for the Headlines dataset. Gold standard labels and scores in rows, system labels and scores in columns. We used a logarithmic scale to produce the color palette.

De Blasio sworn in as NY mayor, succeeding Bloomberg
Bill De Blasio sworn in as NY mayor, succeeding Bloomberg

Regarding the errors in scores, Figure 4.4 shows the confusion matrix on the right, where scores have been rounded to the nearest integer. Most errors are between contiguous scores, with some exceptions like the system returning 4 instead of 2, or 5 instead of 3. This shows a bias of our system towards high scores, which we would like to fix in the future.

Comparison to the state-of-the-art

Table 4.7 shows the results of our best system (FULL) with respect to the state-of-the-art, as set in the SemEval Task 2 competition (Agirre *et al.*, 2015a; Karumuri *et al.*, 2015; Hänig *et al.*, 2015; Bicici, 2015), which included a subtask on Interpretable STS based on our dataset[13]. Our system outperforms the best system (UMDuluth_3) in both datasets, except in the ALI and SCORE results for Headlines.

[13]Participants could send up to three runs. Note that the task also included a track where the gold chunks were made available to participants. For the sake of space, we focus on the most natural track, where systems need to chunk sentences on their own.

	Headlines				Images			
	ALI	TYPE	SCORE	T+S	ALI	TYPE	SCORE	T+S
ExBThemis_a	70.3	43.3	62.2	42.9	69.7	39.7	60.7	38.1
ExBThemis_m	70.3	43.3	62.0	42.9	69.7	39.7	61.1	38.7
ExBThemis_r	70.3	43.3	62.1	42.8	69.7	39.7	60.9	38.7
RTM-DCU	49.1	37.1	45.5	37.1	35.4	22.8	31.9	22.8
SimCompass_c	64.7	43.3	56.4	38.7	54.3	28.5	45.5	24.2
SimCompass_p	63.1	42.8	55.3	38.7	-	-	-	-
SimCompass_w	64.6	43.3	56.2	38.8	54.3	28.3	45.6	24.3
UMDuluth_1	**78.2**	50.6	69.7	50.0	83.4	55.3	75.0	54.3
UMDuluth_2	**78.2**	51.1	69.9	50.5	83.4	57.6	75.1	56.3
UMDuluth_3	**78.2**	51.5	**70.2**	51.0	83.4	56.1	74.6	54.7
FULL	77.1	**53.4**	70.1	**52.2**	**83.9**	**60.9**	**76.1**	**58.8**

Table 4.7: Comparison to the state-of-the-art. Results (F1 %) on each of the datasets. Columns show the results on each evaluation criteria, where T+S stands for "Type and Score". Best results in each column in bold.

The UMDuluth_3 system improved the quality of the publicly available OpenNLP chunker, with some post processing rules, which could explain the better performance of ALI on Headlines. They use the same alignment software as our system. The labeling module is a supervised system based on support vector machines, similar to ours. Our better results can be explained by a larger number of features, which include similarity scores from the scoring module and more WordNet similarity measures. Unlike our system, their scoring module is based on the labels.

The good results of participating systems and the improvement over baselines show that Interpretable STS is a feasible task in all steps: alignment, labeling of relations and scoring similarity. It is also indirect evidence that the task is well designed and the annotation consistent.

Note that we participated in the Semeval task with previous versions of our system. The only difference lays in the successful data-augmentation strategy to train the classifier for alignment labels, which was based on gold standard chunks and now uses a mixture of gold chunks and system-produced chunks (cf. Section 4.4). Overall results (T+S) improved from 47.1 to 52.2

in Headlines and from 56.4 to 58.8 in Images.

4.6 Application of Interpretable STS

In order to judge whether the information returned by an Interpretable STS system can be used to clarify and explain semantic judgments to humans, we performed two user studies. We first devised a verbalization algorithm, which, given two sentences, their similarity score and the typed and scored alignment between chunks, returns English text verbalizing the differences / commonalities between the two sentences. We then contrasted the activities of the users with and without the Interpretable STS verbalizations, trying to show that the verbalizations helped the users in the two case studies.

We decided to verbalize the differences and commonalities using text. Another alternative would be to use a visual interface (e.g. similar to Figure 4.3). One advantage of our verbalization system is that it is applicable in text-only scenarios like chats and also in speech-only scenarios (e.g. using text recognition and text-to-speech synthesis). In an educational setting, students might prefer textual feedback, as they might need to be trained to interpret the graphical interface. In the future, it would be interesting to contrast the effectiveness of each modality in the user cases.

Label	Verbalization produced
EQUI	X and Y mean the same
SPE1	X is [a bit more | more | much more] specific than Y
SPE2	X is [a bit more | more | much more] general than Y
SIMI	X and Y are [very | $\varnothing$ | slightly | scarcely] similar
REL	X and Y don't mean the same but are [closely | $\varnothing$ | somehow | distantly] related
OPPO	X and Y mean the opposite

Table 4.8: Templates employed for producing verbalizations summarized by label. X and Y refer to the aligned chunks from sentence 1 and 2, respectively. The score is used to select the qualifiers in SPE1, SPE2, SIMI, and REL.

Verbalization

Given the output of the Interpretable STS system, we devised a simple template-based algorithm to verbalize the alignment information into natural language. The label of the alignment is used to select which template to use, and the score is used to qualify the strength of the relation, as summarized in Table 4.8. An example of a verbalization for a sentence pair is shown in the bottom of Figure 4.5.

We are aware that the verbalization algorithm could be improved, specially to avoid repetitions, and make the text more fluent and easier to read. It currently produces one sentence per alignment, resulting in too much text. The information from several alignments could be synthesized and summarized in shorter messages. In any case, we will show that this simple verbalization algorithm is effective enough in the two user case studies.

First user study: STS

In the first user study, the volunteers need to score the similarity of the two sentences. Figure 4.5 shows the instructions for the volunteers, which mimic those used to annotate STS datasets (Agirre *et al.*, 2015a). The Figure corresponds to the case where a verbalization is shown to the volunteer. We then measured the agreement of the volunteers with the gold standard STS score. In order to contrast whether the verbalizations had any impact in the performance of the users in the task, we run three scenarios: no verbalization, automatic verbalization based on the Interpretable STS gold standard, automatic verbalization based on the Interpretable STS system output.

Second user study: English students

In the second user study, we consider an English as a Second Language education scenario, where the volunteers play the role of an inspector who is overseeing the grades given by a lecturer to a student. The learning task assigned to each student was to summarize a piece of news into a single headline. The volunteer-inspector is given two sentences: the first one is the

```
Please, evaluate the two sentences with a score between
0 and 5, with the following interpretation:

  (5) The two sentences are completely equivalent, as they mean the same
      thing.
      The bird is bathing in the sink.
      Birdie is washing itself in the water basin.
  (4) The two sentences are mostly equivalent, but some unimportant details
      differ.
      In May 2010, the troops attempted to invade Kabul.
      The US army invaded Kabul on May 7th last year, 2010.
  (3) The two sentences are roughly equivalent, but some important
      information differs/missing.
      John said he is considered a witness but not a suspect.
      "He is not a suspect anymore."
  (2) The two sentences are not equivalent, but share some details.
      They flew out of the nest in groups.
      They flew into the nest together.
  (1) The two sentences are not equivalent, but are on the same topic.
      The woman is playing the violin.
      The young lady enjoys listening to the guitar.
  (0) The two sentences are completely dissimilar.
      John went horse back riding at dawn with a whole group of friends.
      Sunrise at dawn is a magnificent view to take in if you wake up early
      enough for it.

Please, note that you have some explanations below the sentences.
Read them carefully and use them to assign your scores.
-----------------------------------------------------------------
Afghan legislators approve new election law
Afghan president approves new election law

They are very similar.
Note that 'Afghan legislators' and 'Afghan president'
don't mean the same but are closely related
Note also that 'approve' and 'approves' mean the same

[ Write answer ]
```

Figure 4.5: Instructions and task for users participating in the first user study. This example shows a verbalization based on the alignment of the system.

reference headline used by the professor to assess the student, and the second headline is that produced by the student. To simulate the scenario we re-used the pairs of sentences in the Headlines dataset, together with their similarity score. The similarity score simulates the grade given to the student.

The task assigned to the volunteers is thus to assess to what extent they agree with the grading. Users are given the following information: the reference

headline of the professor, the headline done by the student, the grade given, and, optionally, the feedback in the form of the automatically produced verbalization. We collect the feedback (agreement level) in the form of an integer between 0 (complete disagreement) and 10 (complete agreement). Figure 4.6 shows the instructions and one example pair, alongside the grade.

We run three scenarios: no verbalization, automatic verbalization based on the Interpretable STS gold standard, and automatic verbalization based on the Interpretable STS system output.

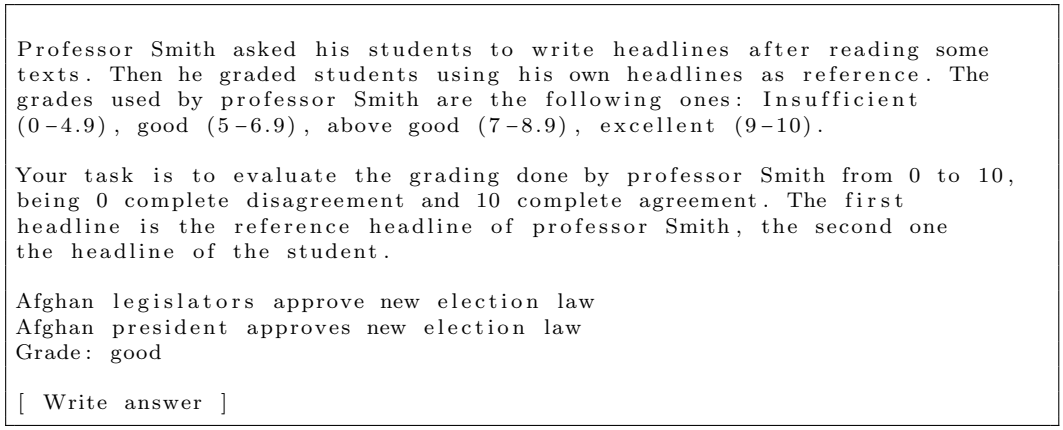


Figure 4.6: Instructions and task for users participating in the second user study. In this example, no verbalization is given to the user.

Setting the task

To conduct the user studies, we randomly selected 48 sentence pairs from the Headlines dataset (see section 4.3). The sentence pairs are accompanied by a gold standard similarity score which ranges from 0 (no similarity) to 5 (equivalence), and we thus sampled the 48 pairs uniformly according to the score. The same set of 48 pairs was used in the two user studies.

The first user study involved 4 native English speakers. For the second user study, which was related to an English as a Second Language setting, we involved 4 non-native English speakers with a verified level C1 of English.

To test whether verbalizations are useful or not, we randomly split the 48

Item sets	No verb	GS verb	SYS verb
A	E1_1	E2_2	E3_2
B	E2_1	E3_3	E4_3
C	E3_1	E4_2	E1_2
D	E4_1	E1_3	E2_3

Table 4.9: Sketch used to distribute item sets (A-D) among participants (E1-E4) with the three possible verbalizations option in the rows. The number after the underscore refers to the order of presentation to the user, e.g. E2_2 is shown to user E2 after E2_1 and before E2_3.

items in 4 item sets (A, B, C and D) and distributed them among participants (E1-E4) according to the sketch shown in table 4.9. The sketch helps organize which files are distributed without verbalizations, which ones are distributed with verbalizations based on gold standard annotations of the Semeval data, and which ones are distributed with verbalizations produced by the system described in section 4.4 (using the system chunk input data). The sketch distributes items across users and verbalizations in a uniform way in order to reduce biases across users, verbalizations and item sets. The same sketch has been used to distribute the files for both scenarios.

Rows from table 4.9 show how each item set with a specific verbalization (No verb, GS verb, SYS verb) is assigned to each participant and in which order. For instance, user E4 will do E4_1, E4_2 and E4_3 in order, that is, the user will first do items in the item set D with no verbalization, then the item set C with GS verbalization and finally the item set B with SYS verbalization. We always show the no verbalized item set first, followed by verbalized itemsets, which are offered in different orders.

Results

To measure the results of the first user study, we use the correlation between the scores given by participants and the gold standard STS score. We follow the tradition on the open evaluation tasks on STS (Agirre *et al.*, 2012, 2015a) and use Pearson coefficient correlation as the main measure, but also report Spearman rank correlation for completeness. Table 4.10 shows the

	No verb	SYS verb	GS verb
Pearson r	83	92	90
Spearman ρ	83	92	91

Table 4.10: First user study: Correlations (%) for non verbalized items, gold standard verbalized items and system verbalized items.

correlation for non-verbalized pairs, gold standard verbalized pairs, and system verbalized pairs. Both correlation measures output similar values, with higher correlation values for the verbalized scenarios, showing that the explanations are indeed helpful in this task. The verbalizations obtained from the system output are comparable to those of the gold standard, showing that approximate performance might be enough for being helpful in this task. We performed significance tests between the verbalization options using Fisher's z-transformation for relatedness (Press *et al.*, 2002, equation 14.5.10). The difference between system verbalization and no verbalizations is statistically significant for both Pearson and Spearman[14], but the p-values for gold standard verbalizations vs. no verbalizations are larger[15]. Finally, the difference between system and gold standard verbalizations is not statistically significant.

In the second user study the results correspond to the agreement level in each scenario (cf. Table 4.11). The table reports both the mean agreement level (the average of the raw agreement level introduced by the user), and the binary agreement (how many times the user entered an agreement of 5 or larger). In this user study, the effect of system verbalizations is not as clear as in the previous case: the binary agreement is better (83 vs. 77) but the mean agreement level is very similar (76 vs. 74). The automatic verbalizations produced using gold standard annotations do have a clear impact in the task (94 vs. 77 binary agreement, and 88 vs. 74 agreement level), as the users tend to agree more with the scores assigned by the lecturer. The difference between system verbalization and no verbalizations is not statistically significant in any case, but the difference between gold verbalizations and no verbalization

[14]p-values of 0.057.

[15]p-values of 0.178 on Pearson and 0.107 on Spearman.

is significant [16].

All in all, the results show that a simple method to produce verbalizations based on Interpretable STS annotations are effective in both user studies, as the users could perform the task better. This is a strong indication that our annotation task is well-defined, and leads to verbalizations which are intelligible and which help the users understand the semantic similarity of the target texts. The results obtained by the Interpretable STS systems are promising, with positive effects in both user studies.

	No verb	SYS verb	GS verb
Agreement level	74	76	88
Binary agreement	77	83	94

Table 4.11: Second user study: Agreement level (%) with grade [0..10] and binary agreement (%) [0..100] with non verbalized items, gold standard verbalized items and system verbalized items.

4.7 Conclusions and future work

This paper presents Interpretable Semantic Textual Similarity, where we formalize an interpretability layer on top of STS. We describe a publicly available dataset of sentence pairs[17], where the relations between segments in each sentence are labeled with a relation type and a similarity score. The labels represent relations between segments such as equivalence, opposition, specificity, similarity and relatedness, together with factuality and polarity differences. The Interpretable STS labels are closely related to those available in Natural Logic or Textual Entailment, and, thus, our dataset is complementary to resources such as those presented in Pavlick *et al.* (2015).

We also present a system for Interpretable STS, based on a pipeline which first identifies the chunks in each input sentence, then aligns the chunks between the two sentences, and finally uses a supervised system to label the

[16] p-values of 0.019 and 0.031 for the agreement level and binary agreement, respectively, using paired t-test.

[17] See Section 4.3 for further details.

alignments and a mixture of several similarity measures to score the alignments. The good results and the improvement over baselines show that Interpretable STS is a feasible task in all steps: alignment, labeling of relations and scoring of similarity. It is also indirect evidence that the task is well designed and the annotation consistent, as supported by the high inter-annotator agreement.

Beyond the evaluation of the annotation layer, we also studied whether the interpretable layer could be useful in final applications. To do so, we constructed a simple verbalization algorithm, which, given two sentences and the Interpretable STS annotations, produces a textual explanation of the differences/similarities between the sentences. We then carried out two successful small-scale user studies, which show evidence that users which had access to the explanations perform the task better. We take this as a preliminary indication that our automatically produced explanations are effective to understand the texts. The interpretability layer defined here is general, and it could be also applied on textual inference datasets (Giampiccolo *et al.*, 2007; Bowman *et al.*, 2015).

In the near future, we would like to improve the performance of the Interpretable STS system. The current system performs each step independently (alignment, labeling and scoring of the chunk pair), but does not enforce consistency. For instance, it can produce a weak relation type like REL and a strong similarity score such as 4.5, or viceversa. In fact, the alignment score could feed the typing, and the type of the alignment could be useful for assigning the score. We are thus currently exploring joint algorithms which would perform some of the steps together, using neural networks as in (Zhou *et al.*, 2016). The error analysis shows that our system has a bias towards equivalence and high scores, which future versions of the system should try to remedy.

We would also like to improve our simple and naive verbalization algorithm, as the effectiveness in real tasks also depends on producing natural-looking text which is up to the point and does not contain superfluous information. As an alternative, in some settings, a graphical interface could be also effective. Finally, we plan to perform a more extensive user study on a real task.

Tutoring systems for English as a second language look like a promising direction for systems that automatically grade students and produce explanations of the grading. Beyond the educational domain, the explanatory layer could be used on other tasks, such as question answering, information extraction or summarization.

Finally, the dataset contains headline and caption texts, which limits the range of linguistic phenomena like opposition, polarity and factuality. Coming up with methods to gather naturally-occurring pairs of sentences exhibiting more linguistic phenomena is an open question. The texts also exhibit simple syntactic structure in both corpora, making them easier than more complex text. Future annotation efforts could focus on more complex sentences like those found in newspaper text, e.g. the Newspaper dataset in (Agirre *et al.*, 2012), or plagiarism datasets (Potthast *et al.*, 2010b).

5. KAPITULUA

Ondorioak eta etorkizuneko lanak

Tesi honetan hizkuntza-ulermenean sakondu dugu, lotutako atazak eta sistemak aztertuz. Konputagailuak hizkuntzaren bitartez komunikatzea erronka handiko ataza da, hizkuntza-ulermenaren maila altua eskatzen duena. Are gehiago, hezkuntzaren domeinuan murgiltzean zailtasunak areagotu egiten dira, ikasleei feedback erabilgarria emateko esaldien **errepresentazio abstraktu zehatzak** behar direlako. Lan honetan hitzen eta hitz n-gramen errepresentazio abstraktuak sortzeko **konposizionaltasuna** eta errepresentazioen arteko **interakzioak modelatzeko** atentzio-mekanismoak ikertu ditugu, baita interakzio horiek anotazio linguistikoen bitartez aberastu ere. Horrekin guztiarekin, gure helburua hezkuntzaren alorreko hizkuntza-ulermenean pauso bat aurrera egitea izan da. Une honetan, hizkuntza-ulermena azkar hazten ari den alorra da, eta aurrekarien egoera abiadura handiz aldatzen ari den arren, hizkuntza-ulermen osoa eta zehatza duen adimen artifizial orokor batetik urrun gaude.

5.1 Hitz n-gramen arteko atentzio-ereduak

Tesiko lehen ikerketa-lerroari erreparatuz (ikus 1.2. Atala) kontribuzio nagusiak **esaldi-mailako HU atazak** ikertzearen ondorio izan dira, honako hauek: antzekotasun semantikoa eta inferentzia logikoa, errepresentazio abstraktuak sortzeko gai diren sistema adimendunak ebaluatzeko erabiliak izan direnak. Ataza horietan sistemek esaldi pareen arteko interakzioak modelatu behar dituzte, emaitza gisa bi esaldien arteko erlazio semantikoa kuantifika-

tzen edo logikoa islatzen duen balio bat itzuliz. Tesi honetan ataza hauek deskribatu, landu eta hezkuntzaren alorrari begira biltzen dituzten gaitasunak eta mugak identifikatu ditugu (ikus 2.5. eta 4.2. Atalak). Gainera, egungo hurbilpenak osatzen dituzten proposamenak plazaratu ditugu. Hain zuzen ere, orain arte esaldi pareen arteko interakzioak modelatzeko erabiltzen diren atentzio-mekanismoak hitz bakanetan oinarrituta egon diren arren; lan honen oinarria, erlazio semantikoak modelatzeko hitzak baino haratago doazen **hitzen n-gramak lerrokatzea** erabilgarria dela da.

Hitz bakanetan oinarritutako atentzio-mekanismoak erabiltzen dituzten sistemak hedatzeko proposamenak bildu ditugu, HU maila areagotzeko asmoz n-grama arbitrarioak lerrokatzeko aukera zabalduz. Horretarako, lehenbizi esaldi pareko n-grama arbitrarioentzat errepresentazio abstraktuak, eta, ondoren, n-grama arbitrario horiek lerrokatuta mantenduko dituen atentzio-matrizea sortu dugu. Atentzio-matrizean n-gramak pisuen bitartez lerrokatuta daude.

Antzekotasun semantikoa eta inferentzia logikoa burutzeko arkitektura mota ugari dagoen heinean, adibidez: esaldi pareak bektore gisa modelatzen dituzten hitz-zakuak, konboluzio-sareak, neurona-sare errepikakorrak eta neurona-sare errekurtsiboak; gure proposamena inplementatzeko aurrerantz elikatzen den neurona-sareez osatutako sistema bat aukeratu dugu: **Decomposable Attention Model** edo *DAM* (Parikh *et al.*, 2016). Sistema honek hainbat abantaila ditu, besteak beste, sinplea, interpretagarria eta eraginkorra izatea, eta, gainera, entrenatzeko baliabide eta denbora gutxi behar izatea.

DAM sistema **n-grama arbitrarioak modelatzeko** egokitu dugu, eta bai jatorrizko DAM sistemarekin konparatuta, baita beste neurona-sare metodoen bitartez hedatutako aldaerekin konparatuta baino emaitza hobeak lortzen dituela ikusi dugu. Horrek, n-grama arbitrarioak modelatzeko eraginkortasuna frogatu du, gainera, atentzio-mekanismoa erregresio gisa ikastean emaitzak hobeak direla ikusi dugu: datu-multzotik datu-multzora 1.6 eta 2.3 zehaztasun puntu artekoa, hurrenez hurren.

Ebaluazioa burutzeko antzekotasun semantikoaren eta inferentzia logikoaren domeinuko bost datu-multzo desberdin erabili ditugu, eta gure proposamenak jatorrizko sistemaren errore-tasa hobetzen du eszenario guztietan, kasu-

rik onenean inferentzia logikoko SICK datu-multzoan % 41eko hobekuntza erakutsiz (11 puntu). Antzekotasun semantikoko SICK eta STS Benchmark datu-multzoetan jatorrizko sistemarekiko lortutako errore-tasaren murrizketa ere aipatzekoa da: % 38koa (8.6 puntu) eta % 29koa (9.4 puntu) hurrenez hurren.

Gure proposamenak errepresentazio abstraktuak modelatzeko baseline sistemekin alderatuta ere emaitza hobeak lortu ditu. Izan ere, segmentazioan oinarritutako gure proposamena konboluzio-sareak eta neurona-sare errepikakorrak zein errekurtsiboak erabiltzea baino eraginkorragoa izan da ebaluatzeko erabilitako datu-multzo guztietan. Hala ere, baseline horiek ere DAM baino emaitza hobeak lortu dituzte. Konboluzio-sareak neurona-sare errepikakorrak baino eraginkorragoak izan direla ikusi dugu, emaitzen arteko alde txikiarekin. Gainera, gure proposamena **entrenatzeko instantzia kopuru mugatuak** dituzten datu-multzoetarako bereziki aproposa dela erakutsi dugu, baita inferentzia logikoko datu-multzo handien **adar zailetarako** ere (*hard splits*). Izan ere, SNLI eta Multi NLI datu-multzoetan hobekuntza txikiagoak lortzen bagenituen ere, instantzia tribialak kenduta errore-tasa asko gutxitzen dela ikusi dugu: % 7.6 (2.2 puntu) eta % 11.1 (4.8 puntu) hain zuzen ere.

Honekin guztiarekin gure hipotesi nagusia frogatu dugu: erlazio semantikoak modelatzeko n-grama arbitrarioak lerrokatzearen eraginkortasuna. N-grama arbitrarioak lerrokatzea hitz bakanak eta esaldien dependentzia guztiak modelatzen dituzten zuhaitz deskonposaketen tarteko pauso gisa ikusten dugu, konputazionalki sortzeko aukera merkea eta azaleko informazio sintaktikoa jasotzeko gai den errepresentazio eraginkor gisa, alegia.

5.2 Esaldien arteko desberdintasunak topatzen eta azaltzen

Aurrekarien egoeran atentzio-mekanismoak eta esaldiko interakzioak modelatzeko mekanismoak geroz eta erabilera zabalagoa dutela ikusita, eta mekanismo horiek zehazten dituzten hitzen arteko interakzioak sistema adimendunek hartzen dituzten **erabakiak azaltzeko giltza** direla ikusita, oinarrizko HU ataza hauen gainean **interpretagarritasuna lantzeko geruza** berri bat gehitu dugu. Horrela, antzekotasun semantikoaren eta inferentzia logikoaren onurak bateratuz adierazgarritasun altuko ataza berri bat definituz: **interpretable semantic textual similarity** edo **iSTS**. Ataza berri horretan esaldietako osagai sintaktikoak (chunkak) linguistikoki motibatuta lerrokatzen dira, esaldi barruko kontzeptuen **antzekotasunak eta desberdintasunak** modu esplizituan adieraziz. Sistema adimendunek maila fineko informazio hau erabili dezakete bai hartutako erabakiak ikasleei azaltzeko, baita beren arrazonamenduaren inguruko **feedbacka emateko**.

iSTS ataza SemEval workshopean bi urtez aktibo egon da, parte-hartze eta interes handia piztuz. Ataza horren antolakuntzarako hiru domeinutako datu-multzoak etiketatu ditugu: berrien ingurukoa bat, irudien deskribapenetan oinarritutakoa bigarrena eta, azkena, ikasleen erantzunez osatutakoa, guztira lerrokatutako 2500 esaldi pare baino gehiago plazaratuz. Datu-multzo horiek parte-hartzaileen sistemak entrenatzeko eta ebaluatzeko atzigarri daude, eta, SemEvaletik kanpora ere aurrekarien egoeran ikerketa bideratzeko erabiliak izan dira. Esaterako, Li eta Srikumar (2016) autoreek antzekotasun semantikoaren emaitzak hobetzeko erabiltzen dituzte aipatutako datu-multzoak, aljebra linealeko metodoak erabilita.

Antzekotasun semantiko interpretagarrian parte hartzeko gai diren bi sistema ere garatu ditugu (Lopez-Gazpio *et al.*, 2016b; Agirre *et al.*, 2015b), bai ikasketa automatikoan oinarritutakoak, baita neurona-sareetan oinarritutakoak ere; eta, aurrekarien egoerako emaitzak lortu dira. Garatutako sistema horietan oinarrituz, eta berbalizazio azpisistema sinple bat erabiliz ikasleei emandako feedbackaren erabilgarritasuna bermatu dugu bi ebaluazio-inguruneetan, **feedbacka laguntza gisa** jasotzean erabiltzaileen emaitzak

hobeak direla erakutsiz.

Bigarren ikerketa-lerro honen helburua **irakaskuntzari dagokion HU** maila areagotzea izan da (ikus 1.2. Atala). Gure ustez, esaldi barruko chunkak zehatz lerrokatuta izateak hezkuntzari lotutako sistementzat, eta bereziki, ikasleei feedbacka eman behar dieten sistema adimendunentzat aurrerapauso handia da. Erabiltzaileekin egindako esperimentuetan antzekotasun semantiko eta inferentzia logiko bateratuaren erabilgarritasuna frogatu dugu, biak batera elkar lanean jarriz esaldien **errepresentazioak ulergarriagoak** izan daitezkeela erakutsiz. Gure helburua antzekotasun semantiko interpretagarriaren bitartez egunera arte ezagututako irakaskuntza atazak pauso bat haratago eramatea izan da. Gainera, ikerketa-lerro honek azkenaldian nagusitzen ari den neurona-sareetan oinarritutako ereduak ulertzeko joerarekin bat egiten du, antzekotasun semantiko interpretagarriak **sistemak behemailan ebaluatzeko** aukera handiagoa ematen baitu.

5.3 Sortutako baliabideak

Azkenik, azpimarratu nahi dugu proposamen berriak plazaratu ez ezik, tesi honek ondoko baliabideak sortu eta komunitate zientifikoan atzigarri utzi dituela: antzekotasun semantikoa, inferentzia logikoa eta antzekotasun semantiko interpretagarria burutzeko gai diren sistemak[1], neurona-sareetan eta ikasketa automatikoan oinarritutakoak; eta antzekotasun semantiko interpretagarriarekin[2] lotutako anotazio-gidalerroak, anotazio-interfazeak eta datu-multzoak.

[1] `https://github.com/lgazpio`
[2] `http://alt.qcri.org/semeval2015/task2/`
[2] `http://alt.qcri.org/semeval2016/task2/`

5.4 Etorkizuneko lanak

Etorkizuneko lanei dagokienez, hurrengo ikerketa-lerroak aipatzea interesgarria iruditzen zaigu:

- 3. kapituluan hitzen n-gramak bektore bidez kodetzeko eta konposizionaltasuna lantzeko gai diren neurona-sareetan oinarritutako sistemak erabiltzen ditugu, kasu guztietan oinarri gisa hitz bakanen hitz-bektoreak erabiltzen dituztenak. Hitz n-gramentzat zuzenean errepresentazio abstraktuak sortzea interesgarria litzatekeela iruditzen zaigu, n-gramaren errepresentazio zehatzagoa dutenak. Bektore berri hauek egitura linguistikoari erreparatuz hizkuntzen egitura konplexua kontuan hartu eta hobeto modelatuko luketen errepresentazio abstraktuak liratekeela uste dugu. Ikerketa-lerro hori "phrasal semantics" gisa ezagutzen da eta azkenaldian indarberritzen ari da. Muturreko adibide bat ematearren argi dago izen+aditz motako konbinazioen errepresentazio abstraktuek hitz bakanen errepresentazio abstraktuen baturak baino esanahi gehiago kodetu behar dutela, esaterako: adarra jo parearen kasurako (Iñurrieta *et al.*, 2017).

- 3. kapituluan deskribatzen dugun sistema entrenatzeko iturri berriak bilatu nahi ditugu. Horretarako, hitzen arteko lerrokatzeetan oinarritutako datu-multzoak erabiltzea egokia izan daitekeela uste dugu. Neurona-sareetan oinarritutako sistemek datu kopuru handiak behar dituzte entrenatzeko eta antzekotasun semantiko interpretagarriaren moduko atazek esfortzu handia behar dute datu-multzoak behe-mailan anotatzeko. Datu-multzoak modu merkeagoan anotatzeko bururatu zaigun proposamen bat itzulpen automatikoko (*machine translation* edo MT) lerrokatze-matrizeak erabiltzea da. Gure ustez antzekotasun semantiko interpretagarrirako datu-multzoak MTko iturrietatik erauzteko aukera dago modu erdi-gainbegiratuan, lerrokatze-matrize hauek sortzeko egin den esfortzua berrerabiliz.

- 4. kapituluan deskribatzen dugun berbalizazio azpisistema hobetu nahi dugu. Horretarako, hizkuntza-eredu neuronalak erabiliz gizakioi naturalagoak litzaizkigukeen esaldiak sortu nahi ditugu. Antzekotasun

semantiko interpretagarriaren bitartez irakaskuntzaren alorrean feedbacka sortzeko lehen urratsak eman baditugu ere, erabilitako berbalizazio azpisistema oso sinplea da eta itxura robotikoa duten esaldiak sortzen ditu. Sortutako esaldien naturaltasuna areagotuz erabiltzaileen balorazioa positiboagoa izatea lortuko genukeela uste dugu.

- 3. eta 4. kapituluetako sistemak hobetzea da beste etorkizuneko lan bat. Gure helburua inplementatutako sistemek logika-erlazio ahulentzat antzekotasun-balio altuak sor ditzaten ekiditea da, eta alderantziz. 4. kapituluan aipatu dugu arazo hori atazak independenteki tratatzearen ondorio dela, eta horrek trinkotasuna galtzera eramaten gaitu. Horregatik, antzekotasun semantikoaren eta inferentzia logikoaren arteko trinkotasuna mantenduko luketen sistemak garatzea da gure etorkizuneko proposamena. Hainbat lanek erakutsi duten moduan antzekotasun-balioen eta inferentzia-kategorien arteko erlazioa badago, ez baitira ataza guztiz independenteak elkarren artean (Vo eta Popescu, 2016). Gure ustez, erlazio hori ustiatzeko gai den sistema batek kategoriak eta balioak independenteki tratatzen dituen sistema batek baino eraginkortasun handiagoa izan behar du. Horretarako, ikasteko galerafuntzio hedatuak (*Multi-cost learning functions*) aztertzea aurreikusten dugu.

Bibliografia

Abney S. Parsing by chunks. *Principle-based parsing: Computation and psycholinguistics. Robert Berwick and Steven Abney and Carol Tenny(eds.)*, 257–278. Springer Science & Business Media, 1991.

Agerri R., Bermudez J., eta Rigau G. Ixa pipeline: Efficient and ready to use multilingual nlp tools. *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*, 26–31, 2014.

Agirre E., Alfonseca E., Hall K., Kravalova J., Paşca M., eta Soroa A. A study on similarity and relatedness using distributional and wordnet-based approaches. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 19–27. Association for Computational Linguistics, 2009.

Agirre E., Banea C., Cardie C., Cer D., Diab M., Gonzalez-Agirre A., Guo W., Lopez-Gazpio I., Maritxalar M., Mihalcea R., Rigau G., Uria L., eta Wiebe J. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Denver, CO, June 2015a. Association for Computational Linguistics.

Agirre E., Banea C., Cardie C., Cer D., Diab M., Gonzalez-Agirre A., Guo W., Mihalcea R., Rigau G., eta Wiebe J. SemEval-2014 Task 10: Multilingual semantic textual similarity. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 81–91, Dublin,

Ireland, August 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/S14-2010.

Agirre E., Cer D., Diab M., eta Gonzalez-Agirre A. Semeval-2012 task 6: A pilot on semantic textual similarity. **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, 385–393, Montréal, Canada, 7-8 June 2012. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/S12-1051.

Agirre E., Gonzalez-Agirre A., Lopez-Gazpio I., Maritxalar M., Rigau G., eta Uria L. Ubc: Cubes for english semantic textual similarity and supervised approaches for interpretable sts. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 178–183, Denver, Colorado, June 2015b. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/S15-2032.

Agirre E., Gonzalez-Agirre A., Lopez-Gazpio I., Maritxalar M., Rigau G., eta Uria L. Semeval-2016 task 2: Interpretable semantic textual similarity. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 512–524, San Diego, California, June 2016. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/S16-1082.

Aleven V., Popescu O., eta Koedinger K.R. Pedagogical content knowledge in a tutorial dialogue system to support self-explanation. *Proceedings of the AIED-2001 Workshop on Tutorial Dialogue Systems*, 59–70, 2001.

Álvez J., Gonzalez-Dios I., eta Rigau G. Validating wordnet meronymy relations using adimen-sumo. *CoRR*, abs/1805.07824, 2018. URL http://arxiv.org/abs/1805.07824.

Angeli G. eta Manning C.D. Naturalli: Natural logic inference for common sense reasoning. *Proceedings of the 2014 Conference on Empirical Methods*

in Natural Language Processing (EMNLP), 534–545. Association for Computational Linguistics, 2014. URL http://www.aclweb.org/anthology/D14-1059.

Ba J. eta Caruana R. Do deep nets really need to be deep? *Advances in neural information processing systems*, 2654–2662, 2014.

Bahdanau D., Cho K., eta Bengio Y. Neural machine translation by jointly learning to align and translate. *In Proceedings of ICLR*, 2015.

Barrn-Cedeo A., Gupta P., eta Rosso P. Methods for cross-language plagiarism detection. *Knowledge-Based Systems*, 50:211 – 217, 2013. ISSN 0950-7051. URL http://www.sciencedirect.com/science/article/pii/S0950705113002001.

Beltagy I., Erk K., eta Mooney R. Probabilistic soft logic for semantic textual similarity. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1210–1219, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P14-1114.

Bengio Y., Simard P., eta Frasconi P. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2): 157–166, 1994.

Bentivogli L., Bernardi R., Marelli M., Menini S., Baroni M., eta Zamparelli R. Sick through the semeval glasses. lesson learned from the evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. *Language Resources and Evaluation*, 50(1):95–124, 2016. ISSN 1574-0218. URL http://dx.doi.org/10.1007/s10579-015-9332-5.

Berant J., Srikumar V., Chen P.C., Vander Linden A., Harding B., Huang B., Clark P., eta Manning C.D. Modeling biological processes for reading comprehension. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1499–1510, Doha, Qatar, October 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D14-1159.

Bergstra J. eta Bengio Y. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305, 2012.

Best C., van der Goot E., Blackler K., Garcia T., eta Horby D. Europe Media Monitor - System description. *EUR Report 22173-En*, Ispra, Italy, 2005.

Bicici E. Rtm-dcu: Predicting semantic similarity with referential translation machines. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 56–63, Denver, Colorado, June 2015. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/S15-2010`.

Bowman S.R., Angeli G., Potts C., eta Manning C.D. A large annotated corpus for learning natural language inference. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 632–642, Lisbon, Portugal, 2015. Association for Computational Linguistics. URL `http://aclweb.org/anthology/D15-1075`.

Brockett C. Aligning the RTE 2006 corpus. *Microsoft Research*, 2007.

Buchanan B.G., Shortliffe E.H., *et al.*. *Rule-based expert systems*, 3 lib. Addison-Wesley Reading, MA, 1984.

Burrows S., Gurevych I., eta Stein B. The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25(1):60–117, 2015.

Burstein J., Tetreault J., eta Madnani N. The e-rater automated essay scoring system. *Handbook of automated essay evaluation: Current applications and new directions*, 55–67, 2013.

Callaway C., Dzikovska M., Matheson C., Moore J., eta Zinn C. Using dialogue to learn math in the leactivemath project. *Proceedings of the ECAI Workshop on Language-Enhanced Educational Technology*, 1–8, 2006.

Cer D., Diab M., Agirre E., Lopez-Gazpio I., eta Specia L. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. *Proceedings of the 11th International Workshop on Semantic*

Evaluation (SemEval-2017), 1–14. Association for Computational Linguistics, 2017. URL http://www.aclweb.org/anthology/S17-2001.

Chang C.C. eta Lin C.J. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3): 27, 2011.

Chen Q., Zhu X., Ling Z.H., Wei S., eta Jiang H. Enhancing and combining sequential and tree lstm for natural language inference. *CoRR*, abs/1609.06038, 2016. URL http://dblp.uni-trier.de/db/journals/corr/corr1609.html#ChenZLWJ16.

Cho K., van Merrienboer B., Bahdanau D., eta Bengio Y. On the properties of neural machine translation: Encoder–decoder approaches. *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, 103–111, Doha, Qatar, October 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W14-4012.

Choi J., Yoo K.M., eta Lee S. Unsupervised learning of task-specific tree structures with tree-lstms. *CoRR*, abs/1707.02786, 2017. URL http://arxiv.org/abs/1707.02786.

Chung J., Gülçehre Ç., Cho K., eta Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555, 2014. URL http://arxiv.org/abs/1412.3555.

Churchland P.S. eta Sejnowski T.J. *The computational brain*. MIT press, 2016.

Collins A.M. eta Quillian M.R. Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior*, 8(2):240–247, 1969.

Collobert R. eta Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning. *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, 160–167, New York, NY, USA, 2008. ISBN 978-1-60558-205-4.

Collobert R., Weston J., Bottou L., Karlen M., Kavukcuoglu K., eta Kuksa P. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537, 2011.

Conneau A., Kiela D., Schwenk H., Barrault L., eta Bordes A. Supervised learning of universal sentence representations from natural language inference data. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 670–680, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/D17-1070`.

Conole G. eta Warburton B. A review of computer-assisted assessment. *ALT-J*, 13(1):17–31, 2005.

Consortium E. Elvira: An environment for creating and using probabilistic graphical models. *Proceedings of the first European workshop on probabilistic graphical models*, 222–230, 2002.

Cooper G.F. NESTOR : a computer-based medical diagnostic aid that integrates causal and probabilistic knowledge. Barne-txostena STAN-CS-84-1031, Stanford University (Stanford,CA US), 1984.

Cybenko G. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Dagan I., Glickman O., eta Magnini B. The pascal recognising textual entailment challenge. *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, 177–190. Springer, 2006.

Davies P. *There's no Confidence in Multiple-Choice Testing.* © Loughborough University, 2002.

Duchi J., Hazan E., eta Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. Barne-txostena UCB/EECS-2010-24, EECS Department, University of California, Berkeley, Mar 2010. URL `http://www2.eecs.berkeley.edu/Pubs/TechRpts/2010/EECS-2010-24.html`.

Dzikovska M., Nielsen R., Brew C., Leacock C., Giampiccolo D., Bentivogli L., Clark P., Dagan I., eta Dang H.T. Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 263–274, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/S13-2045`.

Dzikovska M.O., Bental D., Moore J.D., Steinhauser N.B., Campbell G.E., Farrow E., eta Callaway C.B. Intelligent tutoring with natural language support in the beetle ii system. *European Conference on Technology Enhanced Learning*, 620–625. Springer, 2010a.

Dzikovska M.O., Moore J.D., Steinhauser N., Campbell G., Farrow E., eta Callaway C.B. Beetle ii: a system for tutoring and computational linguistics experimentation. *Proceedings of the ACL 2010 System Demonstrations*, 13–18. Association for Computational Linguistics, 2010b.

Dzikovska M.O., Nielsen R.D., eta Leacock C. The joint student response analysis and recognizing textual entailment challenge: making sense of student responses in educational applications. *Language Resources and Evaluation*, 50(1):67–93, 2016.

Eigen D., Rolfe J.T., Fergus R., eta LeCun Y. Understanding deep architectures using a recursive convolutional network. *CoRR*, abs/1312.1847, 2013. URL `http://arxiv.org/abs/1312.1847`.

Elman J.L. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

Fellbaum C. *WordNet - An Electronic Lexical Database*. MIT Press, 1998.

Finkelstein L., Gabrilovich E., Matias Y., Rivlin E., Solan Z., Wolfman G., eta Ruppin E. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131, 2002.

Fraser A. eta Marcu D. Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, 33(3):293–303, 2007. ISSN 0891-2017. URL http://dx.doi.org/10.1162/coli.2007.33.3.293.

Fu L. Rule generation from neural networks. *IEEE Transactions on Systems, Man, and Cybernetics*, 24(8):1114–1124, 1994.

Fyshe A., Wehbe L., Talukdar P.P., Murphy B., eta Mitchell T.M. A compositional and interpretable semantic space. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 32–41, Denver, Colorado, May–June 2015. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N15-1004.

Ganitkevitch J., Van Durme B., eta Callison-Burch C. PPDB: The paraphrase database. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*, 758–764, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL http://cs.jhu.edu/~ccb/publications/ppdb.pdf.

Giampiccolo D., Magnini B., Dagan I., eta Dolan B. The third pascal recognizing textual entailment challenge. *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, RTE '07, 1–9, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1654536.1654538.

Gittens A., Achlioptas D., eta Mahoney M.W. Skip-gram - zipf + uniform = vector additivity. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 69–76, Vancouver, Canada, 2017. Association for Computational Linguistics. URL http://aclweb.org/anthology/P17-1007.

Goikoetxea J., Agirre E., eta Soroa A. Single or multiple? combining word representations independently learned from text and wordnet. *AAAI*, 2608–2614, 2016.

Gong Y., Luo H., eta Zhang J. Natural language inference over interaction space. *CoRR*, abs/1709.04348, 2017. URL `http://arxiv.org/abs/1709.04348`.

Gonzalez-Agirre A. *Computational models for semantic textual similarity.* Doktoretza-tesia, University of the Basque Country (UPV/EHU), 2017.

Gu J., Wang Z., Kuen J., Ma L., Shahroudy A., Shuai B., Liu T., Wang X., Wang G., Cai J., *et al.*. Recent advances in convolutional neural networks. *Pattern Recognition*, 2017.

Gururangan S., Swayamdipta S., Levy O., Schwartz R., Bowman S., eta Smith N.A. Annotation artifacts in natural language inference data. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/N18-2017`.

Hänig C., Remus R., eta De La Puente X. ExB Themis: Extensive Feature Extraction from Word Alignments for Semantic Textual Similarity. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015: Task 2)*, Denver, CO, June 2015. Association for Computational Linguistics.

Harris Z.S. Distributional structure. *Word*, 10(2-3):146–162, 1954.

Haxby J.V., Gobbini M.I., Furey M.L., Ishai A., Schouten J.L., eta Pietrini P. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001.

He H. eta Lin J. Pairwise word interaction modeling with deep neural networks for semantic similarity measurement. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 937–948, San Diego, California, June 2016. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/N16-1108`.

High R. The era of cognitive systems: An inside look at ibm watson and how it works. *IBM Corporation, Redbooks*, 2012.

Hill F., Reichart R., eta Korhonen A. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41 (4):665–695, 2015. URL `http://www.aclweb.org/anthology/J15-4004`.

Hinton G., Vinyals O., eta Dean J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Hochreiter S. eta Schmidhuber J. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997a.

Hochreiter S. eta Schmidhuber J. Lstm can solve hard long time lag problems. *Advances in neural information processing systems*, 473–479, 1997b.

Iñurrieta U., Aduriz I., de Ilarraza A.D., Labaka G., eta Sarasola K. Rule-based translation of spanish verb-noun combinations into basque. *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, 149–154, 2017.

Jernite Y., Bowman S.R., eta Sontag D. Discourse-based objectives for fast unsupervised sentence representation learning. *arXiv preprint arXiv:1705.00557*, 2017.

Jiang J.J. eta Conrath D.W. Semantic similarity based on corpus statistics and lexical taxonomy. *Proceedings of the 10th Research on Computational Linguistics International Conference*, 19–33. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP), 1997. URL `http://www.aclweb.org/anthology/O97-1002`.

Jordan P., Makatchev M., Pappuswamy U., VanLehn K., eta Albacete P. A natural language tutorial dialogue system for physics. *Proceedings of FLAIRS*, 6 lib., 521–526, 2006.

Jurafsky D. *Speech & language processing*. Pearson Education India, 2000.

Kaiser L., Gomez A.N., Shazeer N., Vaswani A., Parmar N., Jones L., eta Uszkoreit J. One model to learn them all. *arXiv preprint arXiv:1706.05137*, 2017.

Kapashi D. eta Shah P. Answering reading comprehension using memory networks. *CoRR*, 2015.

Karpicke J.D. eta Roediger H.L. The critical importance of retrieval for learning. *science*, 319(5865):966–968, 2008.

Karumuri S., Vuggumudi V.K.R., eta Chitirala S.C.R. UMDuluth-BlueTeam: SVCSTS-A Multilingual and Chunk Level Semantic Similarity System. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015: Task 2)*, Denver, CO, June 2015. Association for Computational Linguistics.

Kim B., Malioutov D.M., eta Varshney K.R. Whi 2016. *Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning*, 2016a.

Kim Y., Jernite Y., Sontag D., eta Rush A.M. Character-aware neural language models. *AAAI*, 2741–2749, 2016b.

Kingma D.P. eta Ba J. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL `http://dblp.uni-trier.de/db/journals/corr/corr1412.html#KingmaB14`.

Kiros R., Zhu Y., Salakhutdinov R.R., Zemel R., Urtasun R., Torralba A., eta Fidler S. Skip-thought vectors. *Advances in neural information processing systems*, 3294–3302, 2015.

Klein D. eta Manning C.D. Accurate unlexicalized parsing. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, 423–430, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. URL `https://doi.org/10.3115/1075096.1075150`.

Knight W. AI's language problem. *MIT review*, 2016. URL https://www.technologyreview.com/s/602094/ais-language-problem/.

Korb K.B. eta Nicholson A.E. *Bayesian Artificial Intelligence, Second Edition*. CRC Press, Inc., Boca Raton, FL, USA, 2nd edition, 2010. ISBN 1439815917, 9781439815915.

Kriesel D. *A brief introduction on neural networks*. Citeseer, 2007.

Lacave C. eta Dez F.J. A review of explanation methods for bayesian networks. *The Knowledge Engineering Review*, 17:107–127, 6 2002. ISSN 1469-8005. URL http://journals.cambridge.org/article_S026988890200019X.

Lan W. eta Xu W. Neural network models for paraphrase identification, semantic textual similarity, natural language inference, and question answering. *arXiv preprint arXiv:1806.04330*, 2018.

Leacock C. eta Chodorow M. Combining local context and wordnet similarity for word sense identification. *WordNet: An electronic lexical database*, 49 (2):265–283, 1998.

Leacock C., Chodorow M., Gamon M., eta Tetreault J. Automated grammatical error detection for language learners. *Synthesis lectures on human language technologies*, 7(1):1–170, 2014.

LeCun Y., Bengio Y., eta Hinton G. Deep learning. *nature*, 521(7553):436, 2015.

LeCun Y., Boser B., Denker J.S., Henderson D., Howard R.E., Hubbard W., eta Jackel L.D. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

Levy O., Zesch T., Dagan I., eta Gurevych I. Recognizing partial textual entailment. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 451–455, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P13-2080.

Lewis M. eta Steedman M. Combined distributional and logical semantics. *Transactions of the Association of Computational Linguistics*, 1:179–192, 2013. URL http://aclweb.org/anthology/Q13-1015.

Li T. eta Srikumar V. Exploiting sentence similarities for better alignments. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2193–2203, Austin, Texas, November 2016. Association for Computational Linguistics. URL https://aclweb.org/anthology/D16-1237.

Liu H. eta Singh P. Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226, October 2004. ISSN 1358-3948. URL http://dx.doi.org/10.1023/B:BTTJ.0000047600.45421.6d.

Liu Y., Liu Z., Chua T.S., eta Sun M. Topical word embeddings. *AAAI*, 2418–2424, 2015.

Lopez-Gazpio I., Maritxalar M., Gonzalez-Agirre A., Rigau G., Uria L., eta Agirre E. Interpretable semantic textual similarity: Finding and explaining differences between sentences. *Knowledge-Based Systems*, 2016a.

Lopez-Gazpio I., Agirre E., eta Maritxalar M. iubc at semeval-2016 task 2: Rnns and lstms for interpretable sts. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 771–776, San Diego, California, June 2016b. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/S16-1119.

Lopez-Gazpio I., Maritxalar M., Gonzalez-Agirre A., Rigau G., Uria L., eta Agirre E. Interpretable semantic textual similarity: Finding and explaining differences between sentences. *Knowledge-Based Systems*, 119:186–199, 2017.

Lukashenko R., Graudina V., eta Grundspenkis J. Computer-based plagiarism detection methods and tools: An overview. *Proceedings of the 2007 International Conference on Computer Systems and Technologies*, CompSysTech '07, 40:1–40:6, New York, NY, USA, 2007. ACM. ISBN 978-954-9641-50-9. URL http://doi.acm.org/10.1145/1330598.1330642.

Luong T., Pham H., eta Manning C.D. Effective approaches to attention-based neural machine translation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Linguistics. URL `http://aclweb.org/anthology/D15-1166`.

MacCartney B. eta Manning C.D. Modeling semantic containment and exclusion in natural language inference. *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, 521–528, Manchester, UK, August 2008. Coling 2008 Organizing Committee. URL `http://www.aclweb.org/anthology/C08-1066`.

Macdonald N., Frase L., Gingrich P., eta Keenan S. The writer's workbench: Computer aids for text analysis. *IEEE Transactions on Communications*, 30(1):105–110, 1982.

Marelli M., Bentivogli L., Baroni M., Bernardi R., Menini S., eta Zamparelli R. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 1–8, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University. URL `http://www.aclweb.org/anthology/S14-2001`.

Melamed I.D. Manual annotation of translational equivalence: The blinker project. *arXiv preprint cmp-lg/9805005*, 1998.

Mikolov T., Chen K., Corrado G., eta Dean J. Efficient estimation of word representations in vector space. *Proceedings of Workshop at the International Conference on Learning Representations (ICLR 2013)*, abs/1301.3781 lib., 2013a. URL `http://arxiv.org/abs/1301.3781`.

Mikolov T., Karafiát M., Burget L., Černockỳ J., eta Khudanpur S. Recurrent neural network based language model. *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

Mikolov T., Sutskever I., Chen K., Corrado G.S., eta Dean J. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 3111–3119, 2013b.

Mohler M., Bunescu R., eta Mihalcea R. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 752–762. Association for Computational Linguistics, 2011.

Munkres J. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38, 1957.

Murdoch W.J., Liu P.J., eta Yu B. Beyond word importance: Contextual decomposition to extract interactions from lstms. *arXiv preprint arXiv:1801.05453*, 2018.

Murphy B., Talukdar P., eta Mitchell T. Selecting corpus-semantic models for neurolinguistic decoding. *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, 114–123. Association for Computational Linguistics, 2012.

Nangia N., Williams A., Lazaridou A., eta Bowman S. The repeval 2017 shared task: Multi-genre natural language inference with sentence representations. *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, 1–10, Copenhagen, Denmark, 2017. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W17-5301`.

Nielsen R.D., Ward W., eta Martin J.H. Recognizing entailment in intelligent tutoring systems. *Natural Language Engineering*, 15(04):479–501, 2009.

Nielsen R.D., Ward W.H., Martin J.H., eta Palmer M. Annotating students' understanding of science concepts. *LREC*, 2008.

Olah C., Satyanarayan A., Johnson I., Carter S., Schubert L., Ye K., eta Mordvintsev A. The building blocks of interpretability. *Distill*, 2018. https://distill.pub/2018/building-blocks.

Pagliardini M., Gupta P., eta Jaggi M. Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. *NAACL 2018 - Conference of the North American Chapter of the Association for Computational Linguistics*, 2018.

Parikh A., Täckström O., Das D., eta Uszkoreit J. A decomposable attention model for natural language inference. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2249–2255, Austin, Texas, November 2016. Association for Computational Linguistics. URL `https://aclweb.org/anthology/D16-1244`.

Pavlick E., Bos J., Nissim M., Beller C., Van Durme B., eta Callison-Burch C. Adding semantics to data-driven paraphrasing. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1512–1522, Beijing, China, July 2015. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P15-1146`.

Pedersen T., Patwardhan S., eta Michelizzi J. Wordnet::similarity: Measuring the relatedness of concepts. *Demonstration Papers at HLT-NAACL 2004*, HLT-NAACL–Demonstrations '04, 38–41, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1614025.1614037`.

Pennington J., Socher R., eta Manning C. Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/D14-1162`.

Potthast M., Barrón-Cedeño A., Eiselt A., Stein B., eta Rosso P. Overview of the 2nd International Competition on Plagiarism Detection. In Braschler

M., Harman D., eta Pianta E., editors, *Working Notes Papers of the CLEF 2010 Evaluation Labs*, September 2010a. ISBN 978-88-904810-2-4. URL http://www.clef-initiative.eu/publication/working-notes.

Potthast M., Hagen M., Göring S., Rosso P., eta Stein B. Towards data submissions for shared tasks: first experiences for the task of text alignment. *Working Notes Papers of the CLEF*, 1613–0073, 2015.

Potthast M., Stein B., Barrón-Cedeño A., eta Rosso P. An evaluation framework for plagiarism detection. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, 997–1005, Stroudsburg, PA, USA, 2010b. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1944566.1944681.

Press W., Teukolsky S., Vetterling W., eta Flannery B. *Numerical Recipes: The Art of Scientific Computing V 2.10 With Linux Or Single-Screen License*. Cambridge University Press, 2002. ISBN 9780521750363. URL http://apps.nrbook.com/c/index.html.

Prijatelj D., Ventura J., eta Kalita J. Neural networks for semantic textual similarity. *International Conference on Natural Language Processing (ICON)*, 2017.

Rajpurkar P., Zhang J., Lopyrev K., eta Liang P. Squad: 100,000+ questions for machine comprehension of text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392. Association for Computational Linguistics, 2016. URL http://www.aclweb.org/anthology/D16-1264.

Rashtchian C., Young P., Hodosh M., eta Hockenmaier J. Collecting image annotations using Amazon's Mechanical Turk. *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT 2010, 139–147, Stroudsburg, PA, USA, 2010. URL http://dl.acm.org/citation.cfm?id=1866696.1866717.

Richardson M., Burges C.J., eta Renshaw E. MCTest: A challenge dataset for the open-domain machine comprehension of text. *Proceedings of the 2013*

Conference on Empirical Methods in Natural Language Processing, 193–203, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D13-1020.

Riordan B., Horbach A., Cahill A., Zesch T., eta Lee C.M. Investigating neural architectures for short answer scoring. *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, 159–168, 2017.

Ritter A., Mausam, eta Etzioni O. A latent dirichlet allocation method for selectional preferences. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 424–434, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P10-1044.

Rumelhart D.E., Hinton G.E., eta Williams R.J. Learning representations by back-propagating errors. *nature*, 323(6088):533, 1986.

Rus V., Lintean M., Moldovan C., Baggett W., Niraula N., eta Morgan B. The SIMILAR corpus: A resource to foster the qualitative understanding of semantic similarity of texts. *Semantic Relations II: Enhancing Resources and Applications, The 8th Language Resources and Evaluation Conference (LREC 2012), May*, 23–25, 2012.

Santos C.D. eta Zadrozny B. Learning character-level representations for part-of-speech tagging. *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 1818–1826, 2014.

Shen T., Zhou T., Long G., Jiang J., Wang S., eta Zhang C. Reinforced self-attention network: a hybrid of hard and soft attention for sequence modeling. *arXiv preprint arXiv:1801.10296*, 2018.

Socher R., Perelygin A., Wu J., Chuang J., Manning C.D., Ng A., eta Potts C. Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642. Association for Computational Linguistics, 2013. URL http://www.aclweb.org/anthology/D13-1170.

Stolcke A. Srilm-an extensible language modeling toolkit. *Seventh international conference on spoken language processing*, 2002.

Suermondt H.J. *Explanation in Bayesian Belief Networks*. Doktoretza-tesia, Stanford University Stanford, Stanford, CA, USA, 1992. UMI Order No. GAX92-21673.

Sultan M.A., Bethard S., eta Sumner T. Back to basics for monolingual alignment: Exploiting word similarity and contextual evidence. *Transactions of the Association for Computational Linguistics*, 2:219–230, 2014.

Sultan M.A., Bethard S., eta Sumner T. Dls@cu: Sentence similarity from word alignment and semantic vector composition. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 148–153, Denver, Colorado, June 2015. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/S15-2027`.

Sultan M.A., Salazar C., eta Sumner T. Fast and easy short answer grading with high accuracy. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1070–1075, 2016.

Sutskever I., Vinyals O., eta Le Q.V. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 3104–3112, 2014.

Tai K.S., Socher R., eta Manning C.D. Improved semantic representations from tree-structured long short-term memory networks. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1556–1566, Beijing, China, July 2015. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P15-1150`.

Tian J., Zhou Z., Lan M., eta Wu Y. Ecnu at semeval-2017 task 1: Leverage kernel-based traditional nlp features and neural networks to build a universal model for multilingual and cross-lingual semantic textual similarity.

Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), 191–197, Vancouver, Canada, 2017. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/S17-2028.

Towell G.G. eta Shavlik J.W. Extracting refined rules from knowledge-based neural networks. *Machine learning*, 13(1):71–101, 1993.

Tversky A. Features of similarity. *Psychological Review*, 84:327–352, 1977.

Upadhyay S., Chang K.W., Taddy M., Kalai A., eta Zou J. Beyond bilingual: Multi-sense word embeddings using multilingual context. *Proceedings of the 2nd Workshop on Representation Learning for NLP*, 101–110, Vancouver, Canada, 2017. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W17-2613.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., eta Polosukhin I. Attention is all you need. *Advances in Neural Information Processing Systems*, 6000–6010, 2017.

Vo N.P.A. eta Popescu O. Corpora for learning the mutual relationship between semantic relatedness and textual entailment. *LREC*, 2016.

Wang X., Liu Y., SUN C., Wang B., eta Wang X. Predicting polarities of tweets by composing word embeddings with long short-term memory. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1343–1353, Beijing, China, July 2015. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P15-1130.

White A.S., Rastogi P., Duh K., eta Van Durme B. Inference is everything: Recasting semantic resources into a unified evaluation framework. *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 996–1005, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. URL http://www.aclweb.org/anthology/I17-1100.

Williams A., Nangia N., eta Bowman S. A broad-coverage challenge corpus for sentence understanding through inference. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1112–1122. Association for Computational Linguistics, June 2018. URL http://www.aclweb.org/anthology/N18-1101.

Wittgenstein L. Investigaciones filosóficas (philosophische untersuchungen). *Londres: Kegan Paul*, 1953.

Yang Y., Yih W.t., eta Meek C. Wikiqa: A challenge dataset for open-domain question answering. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2013–2018. Association for Computational Linguistics, 2015. URL http://www.aclweb.org/anthology/D15-1237.

Yin W., Schütze H., Xiang B., eta Zhou B. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *Transactions of the Association for Computational Linguistics*, 4:259–272, 2016. ISSN 2307-387X. URL https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/831.

Young T., Hazarika D., Poria S., eta Cambria E. Recent trends in deep learning based natural language processing. *arXiv preprint arXiv:1708.02709*, 2017.

Zhao J., Zhu T., eta Lan M. Ecnu: One stone two birds: Ensemble of heterogenous measures for semantic relatedness and textual entailment. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 271–277, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University. URL http://www.aclweb.org/anthology/S14-2044.

Zhao Z., Liu T., Li S., Li B., eta Du X. Ngram2vec: Learning improved word representations from ngram co-occurrence statistics. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*,

244–253, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D17-1023.

Zhou G., Zhou Y., He T., eta Wu W. Learning semantic representation with neural networks for community question answering retrieval. *Knowledge-Based Systems*, 93:75–83, 2016.

Erabilitako terminologia eta laburdurak

Antzekotasun semantiko (*Semantic Textual Similarity, STS*)

Aurrerantz elikatzen den neurona-sare (*FeedForward neural network, FFNet*)

Barne-ebaluazio (*Intrinsic evaluation*)

DA sare (*Deep-Averaging network*)

Hezkuntzaren alorreko HP (*Educational NLP*)

Hizkuntzaren prozesamendu, HP (*Natural Language Processing, NLP*)

Hizkuntza-sorkuntza, HS (*Natural Language Generation, NLG*)

Hizkuntza-ulermena, HU (*Natural Language Understanding, NLU*)

Hitz-zaku (*Bag-of-Word, BoW*)

Hitz-bektore (*Word embedding, word vector, embedding vector*)

Inferentzia logiko (*Natural Language Inference, NLI*)

Internet bidezko ikastaro masibo (*Massive Open Online Course, MOOC*)

Itzulpen automatiko (*Machine translation, MT*)

Kanpo-ebaluazio (*Extrinsic evaluation*)

Konboluzio-sare (*Convolutional Neural Network, CNN*)

Konexio ahul (*Skip connection*)

Konexio zuzen (*Highway connection*)

Metodo distribuzional (*Distributional Semantic Model, DSM*)

Ikasteko galera-funtzio hedatu (*Multi-cost learning function*)

Sare errekurtsibo (*Recursive Neural Network, Tree-Structured Recurrent Network*)

Sare errepikakor (*Recurrent Neural Network, RNN*)

Sare siamdar (*Siamese network*)

Sistema-multzo (*Ensemble model*)