Усё пра апрацоўку натуральнай мовы

Апрацоўка натуральнай мовы, АНМ (па-англійскі: Natural Language Processing, NLP) — гэта машыннае пераўтварэнне вуснай і пісьмовай разнавіднасцяў чалавечай камунікацыі. Методыка, апорай якой з'яўляюцца лінгвістыка і статыстыка ў спалучэнні з машынным навучаннем, імкнецца мадэляваць мову на службе аўтаматызацыі.

Якую карысць можа прыўнесці АНМ у бізнес?

Існуюць мірыяды разнастайных прыкладанняў. Любы вытворчы працэс (або асабістая патрэба), што закранае маўленне або электронны тэкст — памеры, хуткасць або складанасць якіх апынуцца дастатковымі, каб падштурхнуць вас да пошукаў аўтаматызаванай дапамогі — можа пачарпнуць карысць з галіны апрацоўкі натуральнай мовы. Давайце разгледзім сістэмна, што АНМ можа вам прапанаваць. Вось 22 накірункі, праілюстраваныя прыкладамі іх ператварэння ў жыццё і пералікам адпаведных навукова-даследчых ініцыятыў. Мы пачнём з пошуку як другой па даце ўзнікнення тэхналогіі ў галіне аўтаматызацыі кіравання інфармацыяй і закончым аглядам прыкладанняў АНМ, прызначаных для штодзённага, аналітычнага і незвычайнага выкарыстання.

Выманне і пошук інфармацыі

Калі б уся існая інфармацыя захоўвалася ў выразным структураваным выглядзе, нам не прыйшлося б ажыццяўляць яе пошук. Выманне інфармацыі звялося б да здзяйснення запытаў і не больш за тое. Але замест гэтага, абстрактна кажучы, 80 працэнтаў рэлевантнай бізнес-інфармацыі бярэ свой пачатак у масе неструктураваных дадзеных, у асноўным тэкставых. Пераважная большасць тэкстаў з'яўляецца ўвасабленнем «натуральнай мовы» (у супрацьлегласць фармальнай мове, якая выкарыстоўваецца, напрыклад, у камп'ютарным праграмаванні або падчас запісу алгебраічнага ўраўнення). Google, Bing і іншыя пошукавыя сістэмы выкарыстоўваюць АНМ, каб атрымаць тэрміны з тэксту (1), каб запоўніць іх індэксы і для парсінга пошукавых запытаў (2). Гэтыя тэрміны могуць мець «імянную сутнасць»: людзі, назвы кампаній, сімвалы акцый і месцы. Таксама могуць уключацца даты, адрасы, URL, і да т.п.; АНМ дазволіць аўтаматызаваць выманне заканамернасцяў (3) і выманне атрыбутаў, звязаных з тэрмінамі (4) фактычнымі або суб'ектыўнымі: дарагі гадзіннік, чорны аўтамабіль, 4,6 кг рыбы.

Больш прадвінутыя механізмы прымяняюцца ў АНМ для выяўлення адносін (5) («гэта, што»), каб пабудаваць іх графы ведаў. АНМ абслугоўвае механізмы апрацоўкі кампутарызаваных ведаў для такіх кампаній, як Apple Siri, Wolfram Alpha і Google Now, а таксама рэсурсаў для ўласных лексічных аналізаў, такіх як Lexaltyics 'Concept Matrix, пабудаваных з дапамогай АНМ да набора дадзеных Вікіпедыі для ідэнтыфікацыі «тэматычных канцэптаў» і «аспектаў», а таксама танальнасці. Па словах генеральнага дырэктара Lexalytics Джэфа Кэтліна, гэтыя функцыі дазваляюць карыстальнікам лёгка

ствараць класіфікатары для вельмі шырокіх тэм, а таксама згарнуць думкі ў карзіну падабенства. Генератар Сістэматыкі Pingar з'яўляецца яшчэ адным прыстасаваннем той жа ідэі: выкарыстоўваць метады АНМ, каб пабудаваць структуру ведаў з наступным выкарыстоўваннем для пошуку, класіфікацыі і іншых задач, звязаных з упраўленнем інфармацыяй.

Канцэпцыі, тэмы, танальнасць і падабенства, а таксама заўвагі па метадах

«Карзіны падабенства»: гэтыя катэгорыі вызначаюцца аналітыкамі або праз статыстычную кластарызацыю. Класіфікацыя — гэта акт размяшчэння прыкладаў па катэгорыях паводле іх прыкмет, або ў кластары на аснове найбольшай адпаведнасці. Класіфікацыя (6) з'яўляецца часткай задачы АНМ - гаворка ідзе пра групоўку і выразаў, і дакументаў. Адна з варыяцыяй тэрміна «групоўка» ўключае ў сябе стварэнне канцэптуальных класаў, напрыклад «вытворцы транспартных сродкаў» для Fiat, Ford, General Motors, Nissan, Toyota і інш. Іншая варыяцыя уключае крыжаваныя спасылкі — некалькі спосабаў звароту да дадзенай рэчы; у наступным прыкладзе «Барак Х. Абама з'яўляецца 44-м прэзідэнтам Злучаных Штатаў. Яго гісторыя — гэта амерыканская гісторыя... Прэзідэнт Абама нарадзіўся на Гаваях» чалавек называецца чатырма рознымі спосабамі, адзін з іх з дапамогай займенніка («яго»), які адносіцца да пэўнага чалавека толькі ў кантэксце.

Хочаце ўбачыць рэальнае выманне сутнасцяў і выяўленне крыжаваных спасылак? Паспрабуйце дэма сістэму кампутарнай мовы Карпарацыя Цыцэрон. Звярніцеся да вэб-старонкі, дзе я знайшоў гэтыя радкі, http://www.whitehouse.gov/administration/president-obama. Націсніце на адзін з «ён» ці «яго» ў размечаным тэксце, і вы ўбачыце, што гэтыя займеннікі былі правільна суаднесеныя з словазлучэннем «Прэзідэнт Абама».

Мяркую, я таксама прысвою нумары выманню канцэпта (7) і вызначэнню тэмы (8), якія звязаны з выманнем інфармацыі (гл. папярэдні падраздзел) абстрактнага характару. Танальнасць таксама з’яўляецца абстрактнай, хоць аналіз танальнасці (9) можна ахарактарызаваць (у вельмі спрошчаным выглядзе) як іншую задачу класіфікацыі, якая ажыццяўляецца па звычайных катэгорыях (станоўчай, адмоўнай ці нейтральнай), па больш вытанчаных эмацыйных катэгорыях (напрыклад, гнеў, шчасце, сум), або сігналах намераў (напрыклад, купіць, прадаць, абнавіць, адмяніць). Наведайце вэб-сайт экспертаў тэкставай аналітыкі Daedalus для дэма-версіі класіфікатара танальнасці. Онлайн дэмаверсія Nerily дазволіць выцягнуць мноства іншых характарыстык тэкста.

Аналіз танальнасці і пошук меркаванняў з'яўляюцца для мяне цэнтральнымі тэмамі. Я напісаў шмат пра іх, і арганізую два разы на год канферэнцыю, Сімпозіум па аналізу танальнасці (Sentiment Analysis Symposium).

Усё, што тычыцца здабычы інфармацыі, робіць АНМ ключавым актывам для тэкставага аналізу, а таксама мадэлюе і структуруе інфармацыйны склад тэкставых крыніц для бізнес-аналітыкі, аналізу дадзеных, даследавання. (Гэта вызначэнне я напісаў яшчэ ў 2007 годзе, у артыкуле TechWeb, што зараз можна ўбачыць у Вікіпедыі.)

Я адхілюся ад тэмы, каб растлумачыць, што вы можаце аўтаматызаваць чалавечы падыход да шматлікіх задач апрацоўкі натуральнай мовы праз краўдсорсінг, выкарыстоўваючы CrowdFlower для «аналізу танальнасці сіламі чалавека» і іншых сістэм, пабудаваных на платформах, такіх як Amazon Mechanical Turk. Акрамя таго, вы таксама можаце атрымаць танальнасць і іншую інфармацыю, аналізуючы нятэкставыя крыніцы, якія вар'іруюцца ад запісаў аб транзакцыях да малюнкаў і маўлення.

Мы вернемся да мовы трохі пазней. Зараз я распавяду пра апошнюю на гэты момант функцыю, звязаную з класіфікацыяй і падабенствам — гэта распазнаванне плагіяту (10), якая, па вялікім рахунку, уяўляе сабой ацэнку параграфаў вынятага тэксту па прынцыпе падабенства. Апісанне прыкладу з тлумачэннем можна знайсці на сайце канферэнцыі PAN-13, там жа прадстаўлена некаторая інфармацыя і зыходны код у дапамогу Python-праграмістам. PAN расшыфроўваецца з ангельскай як Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection, ці ў перакладзе — аналіз плагіяту, ідэнтыфікацыя аўтарства і выяўленне блізкіх дублікатаў. Я мяркую, PAAINDD выглядае нязграбна ў якасці абрэвіятуры.

Арфаграфія, Граматыка і Стыль

Хочыце пісаць бес памылкаў? На шчасце для вас: АНМ убудаваны у ваша любімае праграмнае забеспячэнне па апрацоўцы тэкстаў. Праверка арфаграфіі (11) — гэта адна з асноўных функцый АНМ. Пры дапамозе функцыі праверкі арфаграфіі праграма выдзяляе слова, якога няма ў слоўніку, і прапануе правільныя варыянты. Калі вы калі-небудзь пісалі ў дакуменце Microsoft Word (або OpenOffice, Google Docs або ў любым з незлічоных іншых асяроддзях аўтарынга), вы бачылі нешта падобнае. Але праверка арфаграфіі не вызначыць дзве памылкі ў фразе «Я пайшоў тады ў ты часы» (ангельскі прыклад I went their at tree o'clock). Паспрабуйце гэтую фразу ў JSpell або ў SpellCheck.net у якасці доказу. Для сінтаксічных памылак вам патрэбна праверка граматыкі. Як машына правярае граматыку?

Лінгвістычны падыход да праверкі граматыкі можа ўключаць вызначэнне часціны мовы з дапамогай стварэння дыяграмы сказа (12), прыклад можна паглядзець тут; разметкі па часцінах мовы (13), глядзіце прыклад у дэмаверсіі універсітэта Іллінойса; або праз вывучэнне сінтаксічных адносін (14), як паказана ў дэмаверсіі Connexor. (Сінтаксічны разбор — гэта адзін з спосабаў, які я разглядаў вышэй у пункце 5 як прыклад выяўлення адносінаў паміж суб'ектамі).

Так якое ж будзе заключэнне некаторых праграм (або сродкаў, гледзячы пра што ідзе гаворка) наконт майго тэксту? Я ўставіў абзац з трох сказаў у адну з іх. Праграма выявіла «3 патэнцыйныя памылкі» - 2 магчымыя памылкі правапісу і абвінавачванне ў шматслоўнасці - і паведаміла, што мой тэкст «невыразны, патрабуе рэдагавання».

(Бясплатная версія не прадугледжвае прадстаўлення дэталяў, таму я апушчу назву праграмы). Паспрабуйце некаторыя іншыя: LanguageTool – праграмнае забеспячэнне з адкрытым зыходным кодам (я не знайшоў яе асабліва карыснай, але вы можаце паспрабаваць) і Stilus, што стварылі мае сябры ў Daedalus.

Трэба ўзгадаць яшчэ два варыянта для стылістычнага аналізу: Lymbix аналізуе танальнасць электроннай пошты праз інструмент ToneCheck, і яшчэ адно цікавае прыкладанне – аўтаматычная мадэрацыя каментароў, хоць я не змог знайсці незалежнага пастаўшчыка, супастаўнага з Adaptive Semantics, які Huffington Post купіў яшчэ ў 2010 годзе.

Рэферыраванне і пераклад

Рэферыраванне тэксту (15) з'яўляецца першай з некалькіх функцый АНМ, пра якія я распавядаю, і ўключае ў сябе разуменне натуральнай мовы (апісана ў пунктах з 1 па 14) і генерацыю натуральнай мовы. Праграма для рэферыравання павінна разумець зыходны тэкст дастаткова, каб згенераваць скарочаную версію, якая не мяняе змест і мэту арыгінала.

У красавіку 1958 года, прагрэсіўны даследчык Ханс Петер Лун у часопісе IBM Journal апублікаваў артыкул, The Automatic Creation of Literature Abstracts, у якім апісаў аўтаматычнае рэферыраванне тэксту: «Статыстычная інфармацыя атрыманая з частаты ўжывання слова і размеркавання слова выкарыстоўваецца машынай, каб вылічыць адносную меру значнасці, спачатку для асобных слоў, а затым для сказаў. Сказы, якія атрымліваюць самую высокую адзнаку ў значэнні, выцягваюцца і выводзяцца, каб трапіць у аўтарэферат».

Распрацоўшчык Андрэас Гор на сваім сайце SplitBrains.org забяспечвае вэб - інтэрфейс для рэсурса Надава Ротэма з адкрытым зыходным кодам Open Text Summarizer. Паспрабуйце!

Машынны пераклад (16) з'яўляецца выдатным прыкладаннем для АНМ. Ён не патрабуе тлумачэння; я проста хачу звярнуць вашу ўвагу на Google Translate, дзе вы можаце паспрабаваць гэта самі. Звярніце ўвагу на функцыю аўтаматычнай ідэнтыфікацыі мовы (17).

Пераклад уключае ў сябе больш, чым проста пераклад слова з адной мовы на іншую. Кожная мова мае свой уласны сінтаксіс і ідыёмы. Перакладчыку, чалавеку ці машыне, трэба зразумець сэнс тэксту і перадаць гэты сэнс на прызначаную мову. Як і рэферыраванне, машынны пераклад прадугледжвае генераванне натуральнай мовы. Гэтаксама і наступны прыклад.

Пытальна-адказныя сістэмы

IBM Watson з'яўляецца найбольш вядомым прыкладам працы пытальна-адказнай сістэмы (17): адказ на пытанне фармулюецца пры дапамозе вымання інфармацыі:

сітуацыйна-адпаведныя факты прымаюць форму, якая адпавядае форме пытання. Калі Watson гуляў у Jeopardy, ён сфармуляваў адказы ў выглядзе пытанняў адпаведна з правіламі гульні; гэта вельмі адрозніваецца ад таго, як ён будзе рэагаваць на лячэбна-дыягнастычныя задачы. Я звярну вашу ўвагу на акадэмічную ілюстрацыю, START, якая была распрацавана Барысам Кацам і яго камандай ў Масачусецкім тэхналагічным інстытуце, а таксама накірую Вас да вэб-сайтаў EasyAsk і Inbenta, якія тлумачаць, як пытальна-адказныя сістэмы могуць працаваць у агульных бізнес-кантэкстах.

Распазнаванне маўлення

Давайце прызнаем, што маўленне таксама з'яўляецца натуральнай мовай, і вызначым распазнанне маўлення (18) і генерацыю, або сінтэз маўлення (19) у якасці яшчэ двух функцый апрацоўкі натуральнай мовы (АНМ).

Маўленне гэта нешта большае, чым проста агучванне тэксту. Яно можа адносіцца да пэўнага жанру, валодаць танальнасцю, настроем, эмацыйнай афарбоўкай, якія можна выявіць у інтанацыйных адценнях слоў і сказаў (пытальны сказ – інтанацыя змяняецца ў канцы) і ў зменах гучнасці маўленчага сігналу, хуткасці і іншых паказчыкаў. Вы зразумееце, што я маю на ўвазе, калі праслухаеце інтэрв'ю з прафесарам універсітэта Рочэстэра Вэндзі Хайнцельманам на падкасце Вучыць Кампутар Чуць Эмоцыі (Teaching Computers to Hear Emotions) электроннай версіі часопіса IEEE Spectrum (часопіс, які выдаецца Інстытутам інжынераў электратэхнікі і электронікі).

Зусім не абавязкова прадстаўляць вуснае маўленне ў выглядзе напісанага тэксту, каб выкарыстоўваць яго для аналізу, і, акрамя таго, транскрыбаванне вуснай мовы (20) упэўнена можа лічыцца адной з задач АНМ. Было праведзена мноства тэарэтычных і практычных даследаванняў фаналагічнага аналізу, які вывучае гукі і гукавыя мадэлі, а таксама існуюць прамысловыя сістэмы, якія ажыццяўляюць галасавы пошук (21) па фанемам і мадэлям. Каб убачыць працэс фанетычнага транскрыбавання ў дзеянні, азнаёмцеся з онлайн дэмаверсіяй, якая дэманструе працу прадукту кампаніі Daedalus.

З іншага боку, сінтэз маўлення па тэксце (22) з'яўляецца асобнай сферай АНМ. Машына агучвае тэкст з правільна размечаным вымаўленнем, інтанацыяй (удакладнена), хуткасцю і г.д. На сайце кампаніі Івона (Ivona) можна паглядзець выдатную дэмаверсію, якая прачытае тэкст на розных мовах і з рознымі акцэнтамі. Дарэчы, кампанія нядаўна ўвайшла ў склад кампаніі Амазон (Amazon), а праграма Івона ўжо выкарыстоўваецца на планшэце Kindle Fire.

Будаўнічыя блокі

Нарэшце, заўвага пра інструменты, пра радкі кода, якімі вы можаце скарыстацца каб сабраць усё ў сваё ўласнае рашэнне, а таксама пра навучанне. Аднак, папярэджваю: я не збіраўся ў гэтым артыкуле сістэматызаваць каталог даступнага праграмнага забеспячэння, паслуг з адкрытым зыходным кодам, або чагосьці іншага.

Кагнітыўны лінгвіст Крыстафер Фіппс адзначае ў сваім Блогу Нікчэмнага Лінгвіста: «На шчасце, сфера АНМ узрасла сярод людзей з дружалюбнымі адносінамі да адкрытага доступу, такім чынам, ёсць шмат рэсурсаў свабодна даступных». Фіппс факусуецца на разуменні тэксту, а для гэтага няма лепшага каталога, чым артыкул на старонцы АНМ Стэнфардскага ўніверсітэта, які завецца «Статыстычная апрацоўка натуральнай мовы і корпусная кампутарная лінгвістыка: анатаваны спіс рэсурсаў», хоць, як папярэджвае Фіппс, гэта не для пачаткоўцаў. Я пералічваю шэраг інструментаў з адкрытым зыходным кодам у леташнім артыкуле "Які найбольш магутны інструмент сэнтымент - аналізу з адкрытым кодам?". Два інструмента я туды не ўключыў, таму што яны не аптымізаваныя для сэнтыменту (тэма артыкула): Apache OpenNLP, і Mallet – інструментар машыннага навучання.

У вашым распараджэнні таксама ёсць мноства сэрвісных прапаноў, якія рэалізуюць АНМ, якія прымяняюцца праз онлайн API, большасць у бясплатным доступе для тэставага або абмежаванага выкарыстання. Наўскід я магу прыгадаць AlchemyAPI, Apicultur, Bitext, Clarabridge, OpenAmplify, Pingar, Saplo, Semantria (пры падтрымцы the Lexalytics Salience engine), і Viralheat. Mashape пералічвае многія іншыя. Магчымасці, якасць і кошт вар'іруюцца ў шырокіх межах. Некаторыя робяць толькі маркіроўку сэнтымента або сутнасці, а іншыя аналізуюць больш элементаў тэкста. The Apicultur service і Веб API для Python NLTK Якоба Перкінса, на сайце text-processing.com, з'яўляюцца прыкладамі элементнага аналізу тэкста. Я апушчу дэталі, але, можа быць, калі-небудзь напішу іх у артыкуле, і я таксама не збіраюся пісаць зараз пра праграмы, якія вы можаце ўсталяваць самі, бо іх не так лёгка проста "паспрабаваць", як вэб API.

Што тычыцца навучання, акрамя самастойнага навучання, што можа быць лепш, чым онлайн-курс? Адзін з гэтых курсаў ёсць на Coursera, пад кіраўніцтвам Майкла Колінза з Калумбійскага ўніверсітэта, а таксама ў онлайн-рэжыме можна азнаёміцца з відэа і лекцыйнымі матэрыяламі з папулярнага курсу Стэнфардскага ўніверсітэта Крыстафера Мэнінга. Трэці варыянт — гэта вядомы курс Statistics.com пад кіраўніцтвам доктара Ніціна Індуркья.

Цяпер вы ведаеце

У бізнес-кантэкстах, вы часцей за ўсё будзеце ўжываць АНМ у спалучэнні са сборам, інтэграцыяй і аналізам разнастайных формаў онлайн, сацыяльных і карпаратыўных дадзеных. Усё гэта харaство вымання тэксту, якое я апісваў: у сучасным свеце неаднастайных вялікіх даных гэта не можа быць само па сабе. Гэта зацвярджэнне слушнае для бізнес-аналітыкі - вы атрымліваеце ўздым, ужываючы і інтэгруючы адпаведную разнастайнасць метадаў і дадзеных, - і гэта таксама дакладна і для дзейнасці, якая, здаецца, зусім не звязана з нятэкставымі або немаўленчымі крыніцамі, такой, як вэб-пошук. Нават у тых апошніх выпадках, разумныя сістэмы (sense-making engines) ўлічваюць твой профіль, месцазнаходжанне, мінулую актыўнасць онлайн і ў сацыяльных сетках, сацыяльныя сувязі ў спалучэнні з АНМ, каб забяспечыць лепшыя сітуацыйна-рэлевантныя вынікі.

Апрацоўка натуральнай мовы можа быць вельмі карыснай для вас. Гэта неабходны інструмент для перадавой аналітыкі. Разуменне гэта толькі пачатак.

Аўтар арыгінальнага тэкста — **Сэт Граймз**

Перакладзена з англійскай мовы **Юліяй Луконінай**, **Юліяй Барадзіной** і **Аленай Скопінавай**