WebBrain: Spojené neurální učení se rozsáhlým obecným znalostem světa tzv. common sense knowledge

Tento abstrakt slouží k představení práce nesoucí název WebBrain: Joint Neural Learning of Large-Scale Commonsense Kowledge [1], jež byla prezentována na mezinárodní konferenci ISWC 2016 a následně publikována jako příspěvek ve sborníku konference. Jejími autory jsou Jiaqiang Chen, Niket Tandon, Charles Darwis Hariman a Gerald de Melo.

Za poslední desetiletí se stalo obrovské množství strojově čitelných znalostí dostupné jak prostřednictvím rozsáhlých znalostních grafů jako DBpedie, YAGO či GeoNames, tak skrz rozšířenou adopci standardů jako schema.org. Pro získání dalších znalostí navíc existují techniky umožňující extrahovat informace z prostého textu. Využití těchto znalostí bylo doposud primárně zaměřeno na vylepšení výměny informací a jejich integrace za účelem dosažení lepších výsledků webových vyhledávačů. Je tak zatím nedostatečně využit jejich potenciál pro realizaci inteligentních systémů, které by dokázaly pružně vytvářet pokročilé závěry z informací umístěných na Webu.

Souhrnně lze říci, že výše uvedený příspěvek se zabývá představením nového přístupu k získávání tzv. commonsense knowledge čili obecných znalostí světa, které lze také chápat jako obecné porozumění pojmům a objektům, se kterými se lze setkat v běžném každodenním životě. Cílem, jenž si kladou autoři příspěvku, je realizace takového systému, který by byl schopen správně posuzovat pravdivost faktů týkajících se obecných znalostí světa a zároveň by byl založen na znalostech umístěných na Webu. Obstarávání obecných znalostí světa však není triviálním úkolem, jelikož na Webu tyto znalosti vždy nejsou dostupné přímo, protože je považováno za samozřejmé, že jsou s nimi všichni obeznámeni a bylo by zvláštní, či dokonce nepatřičné, je tedy uvádět. Tento nový přístup se tak zabývá získáváním obecných znalostí světa jak z explicitních informací, tak z implicitních textových informací. Explicitní informace jsou pak samotné trojice tzv. triplety a implicitní informace se týkají souběžného výskytu slov v rozsáhlém množství dat na Webu. Příspěvek dále také dokumentuje experimenty, které byly provedeny pro ověření schopnosti v něm uvedených metod dosáhnout vytyčeného cíle. Výsledkem těchto experimentů je potvrzení, že zvolené metody jsou schopné využít obecnou sémantiku pro učení se tzv. commonsense souvislostí, což například znamená, že přístup umožňuje odvodit z informace „orli umí létat“, že jestřábi pravděpodobně také umí létat. Experimenty také ukazují, že tento přístup je schopen současně získat a pracovat s konceptuálními (pojmovými) vazbami a významy slov.

Následující část tohoto abstraktu se zaměřuje na stručnou prezentaci metod, jež byly využity v tomto novém přístupu pro dosažení vytyčeného cíle.

Získání obecných znalostí světa, tedy commonsense knowledge, bylo předmětem studií po mnoho let. Za zmínku zde stojí dva slibné směry - projekt ConceptNet využívající tzv. crowdsourcing a rozsáhlé dolování dat tzv. data mining spojené s metodou bootstrapping.

Nehledě však na zvolený směr, vždy je nutné zobecňovat a rozšiřovat nad rámec toho, co bylo explicitně získáno, což vede k úkolu doplňování znalostní báze. V kontextu strojového učení se tento úkol obvykle týká schopnosti predikce vztahů čili vazeb. S použitím trénovacího setu, jenž obsahuje reálné příklady tripletů, je tak cílem získat naučený model, který pak může být použit k rozhodování, zda je nový dříve neviděný triplet pravdivý, či nikoli. Zatímco vazba sama se v trénovacím setu objevila, konkrétní triplet nikoli (viz výše uvedený příklad o schopnosti létat u různých druhů ptactva). K úkolu doplnění znalostní báze pak lze přistoupit dvěma způsoby – buď pomocí tenzorové či maticové metody, nebo pomocí neurální sítě pro predikci vazeb, kde nejznámější sítí tohoto druhu je tzv. TransE přístup.

V tomto kontextu příspěvek navrhuje tzv. joint model, jenž se od těchto dvou přístupů liší tím, že se učí zobecňovat nejen na základě existujících tripletů, jako v případě tenzorové a matrixové metody či neurálních modelů založených na TransE přístupu, ale navíc využívá sémantické informace odvozené z rozsáhlých textových dat. Joint model dále vychází z tzv. word2vec CBOW přístupu, který umožňuje lepší škálovatelnost a rychlejší optimalizaci oproti jeho předchůdcům.

Co se týče samotného získání obecných znalostí světa, je všeobecné známé, že lze využít shody faktů pro vytvoření vzorů a ty pak použít k dolování nových znalostí. Bohužel tento přístup tzv. bootstrapping trpí při aplikaci na rozsáhlá data na Webu výrazným šumem, rozsáhlá data jsou však esenciální pro získání dostatečného množství trénovacích dat. Problém je ještě zhoršen samotným cílením na obecné znalosti světa, které typicky nejsou vyjádřeny explicitně. Z tohoto důvodu přístup, prezentovaný v tomto příspěvku, zavádí upravenou metodu bootstrapping navrženou tak, aby minimalizovala šum při aplikaci na rozsáhlá data na Webu. V rámci této metody se pak přístup řídí pravidlem, že dobrý vzor by se měl shodovat s mnoha fakty, ale neměl by být příliš obecný. Vybrané vzory se pak použijí k nalezení nových výskytů v rozsáhlých datech na Webu. Pro lepší porozumění je uveden následující příklad: vzoru <x> je <y> by mohlo kupříkladu odpovídat "slunce je jasné", tím získáme slova "slunce" a "jasné" jako tvrzení pro vazbu máVlastnost. Pro odfiltrování šumu z tvrzení je nutné zkontrolovat, zda získaná slova odpovídají specifikaci požadované domény a rozsahu pro vazbu, k čemuž se používá Word-Net hypernymická taxonomie čili taxonomie všeobecnějších významů.

Po získání tripletů z textu je, jak již bylo uvedeno dříve, dalším krokem trénování modelu pro učení se reprezentace commonsense slov a vazeb, tedy doplňování znalostní báze. Joint model navržený v tomto příspěvku bere extrahované znalosti a zobecňuje je pomocí současného využití nejen extrahovaných dat, ale také rozsáhlého množství generického textu. Konkrétně to například znamená, že bylo o orlech zjištěno, že létají, avšak to samé nebylo zjištěno o jestřábech. Zde se pak model snaží odvozovat na základě jak extrahovaných faktů, tak obecné sémantické příbuznosti mezi orly a jestřáby, při čemž využívá souběžného výskytu slov v rozsáhlém množství textu. Dále tento model umožňuje učit se reprezentaci slov nejen z obecných kontextových informací, ale také z extrahovaných vazeb, kde aplikuje princip říkající, že jeden by měl znát slovo podle společnosti, kterou si u sebe drží. Esenciální je však každopádně společné využití jak obecných kontextových informací, tak vydolovaných vazeb.

Všechny zvolené metody využívané představovaným joint modelem jsou detailně zdokumentovány a vysvětleny v samotném příspěvku. Dále je zdokumentováno také jejich využití v praxi v sekci, jež pojednává o provedených experimentech, pro prokázání funkčnosti nabízeného přístupu. Experimenty jsou prováděny dva – první, využívající dump anglické Wikipedie a Google Web 1T N-Gram dataset, a druhý provedený nad DBpedií.

Příspěvek tedy nabízí komplexní představení WebBrain, nového přístupu k získání a modelování znalostí, který činí další krok k cíli vybavit počítače obecnou znalostí světa tzv. commonsense knowledge, a tak umožnit realizaci inteligentnějších systémů. Ačkoli jsou výsledky tohoto přístupu stále daleko od skutečného commonsense porozumění a inteligence, vytváří model schopný naučit se vektorovou reprezentaci slov a vazeb, která odráží, jak jejich obecnou sémantiku, tak základní commonsense fakta o světě, čímž umožňuje nalézt správnou odpověď i na něco, co nebylo přímo pozorováno v textu.

Bibliografie

[1] CHEN, Jiaqiang et al. *WebBrain: Joint Neural Learning of Large-Scale Commonsense Knowledge*. In: . s. 102. DOI: 10.1007/978-3-319-46523-4_7. Dostupné také z: http://link.springer.com/10.1007/978-3-319-46523-4_7