Relation extraction methods for biomedical literature

Bui, Q.C.

Publication date
2012

Samenvatting

Methodes voor het extraheren van relaties in biomedische teksten spelen een cruciale rol in het automatisch vergaren van feiten en bevindingen die nodig zijn voor levenswetenschappen. Hoewel vele mogelijke benaderingen zijn gepresenteerd in de huidige literatuur blijft het extraheren van relaties een groot probleem. Dit komt onder andere door de kwaliteit van de relaties, benodigde computerkracht, en de verschillende types van relaties die mogelijk zijn. Ten eerste moet de 'precisie' van extractiesystemen worden bevorderd om te voldoen aan de standaarden voor bijvoorbeeld het bouwen van databases met biologische informatie. Ten tweede kosten de meeste bestaande extractiesystemen teveel computerkracht waardoor ze niet kunnen worden toegepast op de literatuur op grote schaal. Hierdoor zijn de huidige systemen nog niet klaar voor real-time toepassingen. Ten derde richten de meeste bestaande methoden zich specifiek op interacties tussen proteines (PPI); vele andere types van relaties zijn nog niet of nauwelijks bestudeerd.

In mijn dissertatie lever ik een bijdrage aan dit onderzoeksgebied door het bestuderen van drie verschillende relatie-extractietaken onder invloed van o.a. de beschikbaarheid van trainingsdata en de verschillende types van relaties. Door het bestuderen van de methodes om deze types van relaties te extraheren kunnen we meer begrijpen over de eigenschappen die zij gemeen hebben. Dit helpt bij het ontwikkelen van generieke extractiemethodes die toepasbaar zijn op meerdere relatietypes. Door het variëren van de beschikbaarheid van data om de modellen mee te kalibreren kunnen we beter begrijpen hoe dit de prestaties van extractiemethodes beïnvloedt. In het bijzonder zijn mijn meest belangrijke resultaten als volgt.

We introduceren een nieuwe methode voor het extraheren en combineren van relaties tussen HIV-medicijnen en mutaties in de genetica van het virus. Onze extractiemethode is gebaseerd op 'natural language processing' (NLP) dat grammaticale relaties produceert en een lijst van regels toepast op deze relaties. Uit onze resultaten blijkt dat ons systeem goed presteert in vergelijking met andere prominente systemen. Ons systeem wordt toegepast in het ViroLab project (www.ViroLab.org) om virologen te helpen bij het vinden van relaties tussen HIV-medicijnen en de mogelijke mutaties van het HIV-virus in een gecontroleerde en geautomatiseerde manier.

We presenteren een hybride systeem dat gebaseerd is op zowel het toepassen van een vaste lijst van regels als het gebruik van kunstmatige intelligentie ('machine learning') om PPIs te extraheren uit de literatuur. We introduceren een strategie voor het opdelen van de gegevens die de complexiteit van de kunstmatige-intelligentie-algoritme significant vermindert, hetgeen de robuustheid van het systeem bevordert. We demonstreren dat ons systeem vergelijkbaar presteert met de state-of-the-art systemen met tienvoudige kruisvalidatie. Bovendien scoort ons systeem beter onder

het ‘cross corpora’ criterium, waar de trainingsdata en de testdata hele verschillende eigenschappen kunnen hebben, hetgeen meer lijkt op de realiteit. Verder vergt ons systeem de minste computerkracht.

We presenteren een nieuwe op regels gebaseerde methode voor het extraheren van biomedische gebeurtenissen uit tekst die bestaat uit twee fases: een leerfase en een extractiefase. Door het gebruik van een gestructureerde representatie delen we de complexe en geneste structuren van biomedische gebeurtenissen op in eenvoudigere syntactische lagen. Op basis van deze structuur leert het systeem nieuwe regels uit de trainingsdata en past de regels ook toe om gebeurtenissen te extraheren. Deze representatie stelt ons niet alleen in staat om de leerfase en de extractiefase te vereenvoudigen, maar vergt ook minder syntactische analyse van de zinnen in de invoertekst. De resultaten van de evaluatie laten zien dat ons systeem goed presteert zowel op samenvattingen als op data bestaande uit volledige teksten van artikelen. Bovendien behaalt het de beste computationele efficiëntie: het is 150 tot 200 keer sneller dan de huidige state-of-the-art systemen. Verder is onze benadering simpel en generiek en kan daarom gemakkelijk worden aangepast voor het extraheren van elk ander type van relaties.