Distributional Semantics for Entity Relatedness

Nitish Aggarwal, Kartik Asooja, Paul Buitelaar
Insight Centre for Data Analytics,
National University of Ireland
Galway, Ireland
firstname.lastname@insight-centre.org

Abstract

Wikipedia provides an enormous amount of background knowledge to reason about the semantic relatedness between two entities. In this work, we present a distributional semantics based approach for computing entity relatedness, and a focused related entities explorer based on this approach.

1 Introduction

Entities like persons, locations, organizations etc. are the key features to define the semantics of natural language text. Significance of measuring relatedness between entities has been shown in various tasks which deal with information retrieval (IR), natural language processing (NLP), text analysis or other related fields. Today, major search engines suggest related entities to the users' queries, which provides the users an opportunity to explore more and extend their knowledge. Recent statistics on search queries suggest that more than 40% of the search queries revolve around a single entity. Google and Yahoo! recommend the related persons, locations, organizations, movies, songs and events by using their knowledge graphs. The publicly available structured knowledge resources such as DBpedia and Freebase consist of limited types of relations which can be defined between two entities, consequently missing many connections which might appear in the real world. Many such missing connections can be explored through unstructured knowledge sources like news articles or Wikipedia articles. Moreover, the structured knowledge resources do not provide any quantification of the relationship strength, which makes it hard for humans to easily explore and traverse the relationships. Reasoning about the semantic relatedness of "apple" and "next" requires an immense amount of world knowledge about the concepts represented by these two surface forms. The semantic meaning of "apple" may refer to a fruit, persons surname or a company. Similarly, "next" refers to more than 20 different entities on Wikipedia but the most common meaning of "next" that comes first in mind is "succeeding item". It is hard to assess the relatedness between "apple" and "next" as they are highly ambiguous. However, if we are given that both apple[1] and "next"[2] are software companies founded by Steve Jobs, we can easily judge their appropriate relatedness. In this work, we briefly explain our distributional semantics model (DSM) for computing entity relatedness dubbed as DiSER (Aggarwal and Buitelaar, 2014), and Entity Relatedness Graph (EnRG)[3] which is a related entities explorer over Wikipedia entities (Aggarwal et al., 2014).

2 DiSER

Semantic meaning of an entity can be inferred from its distribution in a high dimensional space of concepts derived from Wikipedia. We develop an approach called Wikipedia-based Distributional Semantics

[1] http://en.wikipedia.org/wiki/Apple Inc.
[2] http://en.wikipedia.org/wiki/NeXT
[3] EnRG Link: http://enrg.insight-centre.org/

for Entity Relatedness (DiSER), which builds the semantic profile of an entity by using Wikipedia concepts. It generates a high dimensional vector by taking every Wikipedia article's topic as a vector dimension, and associativity weight of an entity with the topic as the magnitude of the corresponding dimension (Aggarwal and Buitelaar, 2014). To measure the semantic relatedness between two entities, it computes the cosine score between their corresponding DiSER vectors. DiSER considers only the hyperlinks in Wikipedia, thus keeping all the canonical entities that appear with hyperlinks in Wikipedia articles. For instance, there is an entity *e*, DiSER builds a semantic vector *v*, where $v = \sum_{i=0}^{N} a_i * c_i$ and c_i is i^{th} concept in the Wikipedia concept space, and a_i is the tf-idf weight of the entity e with the concept c_i. Here, N represents the total number of Wikipedia concepts. DiSER has been shown to outperform state of the art methods and achieves a significant improvement in entity ranking and disambiguation tasks over other methods (see for more details (Aggarwal and Buitelaar, 2014)). Moreover, context around the entity in the text can be used to generate DiSER vectors for entities not having a Wikipedia page.

3 EnRG

EnRG (Aggarwal et al., 2014) is a focused related entities explorer over Wikipedia[4] entities, which utilizes DiSER approach for computing the entity relatedness. EnRG provides the users with a dynamic set of filters and facets for exploring the related entities with the help of DBpedia. Every Wikipedia article has a corresponding DBpedia page that provides further exploration for different relations of the entity. DBpedia defines rdf:type[5] of every Wikipedia entity, which allows us to categorize the ranked lists. These types include classes mainly from DBpedia ontology[6] and YAGO[7]. DBpedia ontology covers abstract types like Person, Company, Location, Movie and others, while YAGO also provides very specific types like American film actors, People from Manhattan, etc. This information enable us to group the related entities under different types. The application also presents two types of provenance information, one based on Wikipedia giving the Wikipedia articles which contain both the related entities, while the other performs a Google web search giving the provenance information from the web. It also shows the relatedness strength quantifying the entity relatedness. The application also gives query suggestions referring to related entities based on the searched query string.

4 Conclusion and Future Work

We presented DiSER, a DSM based approach for computing entity relatedness, which outperforms state of the art methods. DiSER is based on Wikipedia, as it consists of world knowledge about millions of entities. We also discussed a web application EnRG based on DiSER, which can be used to explore related entities. As future work, we aim to apply DiSER and EnRG in various tasks like information retrieval (semantic search), text mining and knowledge base population. Moreover, it would be interesting to extend the coverage of entities in other languages too by using multilingual Wikipedia.

References

Aggarwal, N., K. Asooja, P. Buitelaar, and G. Vulcu (2014). Is brad pitt related to backstreet boys? exploring related entities. *Semantic Web Challenge at the International Semantic Web Conference*.

Aggarwal, N., K. Asooja, H. Ziad, and P. Buitelaar (2014). Who are the american vegans related to brad pitt? exploring related entities. In *24th International World Wide Web Conference (WWW 2015), Florence, Italy*.

Aggarwal, N. and P. Buitelaar (2014). Wikipedia-based distributional semantics for entity relatedness. In *2014 AAAI Fall Symposium Series*.

[4]Snapshot of English Wikipedia from October 2013.

[5]http://www.w3.org/1999/02/22-rdf-syntax-ns#type

[6]http://wiki.dbpedia.org/Ontology

[7]www.mpi-inf.mpg.de/yago/