Semantic Tagging with Linked Open Data

John Cuzzola, Zoran Jeremic, Ebrahim Bagheri
Ryerson University

Dragan Gasevic
Athabasca University

Jelena Jovanovic
University of Belgrade

Reza Bashash
SideBuy Technologies

***Abstract*—Making sense of text is a challenge for computers particularly with the ambiguity associated with language. Various annotators continue to be developed using a variety of techniques in order to provide context to text. In this paper, we describe Denote – our annotator that uses a structured ontology, machine learning, and statistical analysis to perform tagging and topic discovery. A short screencast for the curious is also available at http://youtu.be/espItTRQVzY as well as demonstration links provided in the conclusion.**

Keywords—semantic web, disambiguation, entity recognition, annotators, tagging, wikifying, linked-data, LOD

I. Introduction

The availability of structured link open data, through initiatives such as the "Linked Open Data (LOD)" project[1], has given rise to a new class of annotators for unstructured text. Annotators like TagME [1], DBPedia Spotlight [2], and Alchemy[2] all offer such capability. In this systems paper we describe Denote – our semantic tagging platform based on Linked Open Data. In section II, we outline Denote's algorithm, describe its vocabulary, and key features. In III, we demonstrate these features and compare Denote's output with other annotators.

II. Denote's Design

Denote searches its ontology for similar concepts to the input text by performing keyword extraction then calculating a weighted Jaacard coefficient on resource descriptions. This provides a measure of text similarity. For each resource, its known categories (defined in the ontology) are subjected to a Bayesian filter to exclude those resources and categories that do not appear relevant. This provides a measure of semantic similarity. The surviving resources are then used for the annotations. Denote's output is in the form of a synopsis whose lexicon is given in Table I. The output is a single sentence per annotation with a set of relevant URIs sorted in order of likelihood with confidence and available support statistics.

"Text" [Is_A {}] [[[With_Value •] Of_Units •] |
Acting_As {}] [Cat_Of {}]

Fig. 1. The output of an annotated text.

Denote uses a database of linked open data, represented in the form of n-triples (<subject><predicate><object>), to perform annotations, similarity identification, disambiguation and topic categorization. Denote's database is DBPedia [3]; an ontology derived from Wikipedia. In this respect, it resembles DBPedia Spotlight (DBPedia) and TagME (Wikipedia). However, Denote distinguishes itself in key ways. First, it attempts to assign context to the annotations by its [Acting_As] lexicon. Second, it attempts to annotate numbers [With_Value] through statistical analysis of similar concepts whose <predicate>:<object> are of the same data type [Of_Units]. Third, Denote has an extensive list of topic categories, made available through DBPedia's <dcterms:subject> predicate, which it assigns to its annotations [Cat_Of]. These key differences were the motivation for Denote's creation. While other annotators perform in a similar manner by first spotting word phrases and linking them to the disambiguated top-surface form;- Denote attempts to find related concepts that will be used to determine the properties of the spotted word phrases. This allows for role-based annotations [Acting_As]. We coin this process as *deep tagging* as opposed to the *shallow tagging* of Denote's peers.

TABLE I. Denote's Annotation Lexicon Explained

Lexicon	Explanation
Is_A {}	"is a", "is an", "is used by". Asks: What is it?
Acting_As {}	Context/role. Asks: How is it used?
With_Value •	If number, Asks: What is the number value?
Of_Units •	If number, Asks: What is the units of measure?
Cat_Of {}	Asks: What relevant topic categories?

III. Demonstration

In this section, we describe three core functions in Denote's toolkit: text annotation, number annotation, and category disambiguation.

A. *The Text Annotator*

Table II demonstrate Denote's capabilities when compared to TagME and DBPedia Spotlight using the same input text of: "*BLT. The sub that proves great things come in threes. In this case, those three things happen to be crisp bacon, lettuce and juicy tomato. While there's no scientific way of proving it, this BLT might be the most perfect BLT sandwich in existence*. The default configuration for Denote, TagME and Spotlight were unchanged. Spotlight does not perform category analysis. TagMe gives a topic listing but this list is simply the annotated text rather than a separate categorization. Consequently, the [Cat_Of] portion of Denote's synopsis was omitted and left for part C.

DBPedia Spotlight was the least effective with the fewest annotations and an incorrect disambiguation of BLT as a "Bizarre Love Triangle". TagME performed well with

[1] http://linkeddata.org/
[2] http://www.alchemyapi.com/

numerous annotations with few mistakes (incorrectly tagged words “crisp” and “juicy”. Both Denote and TagME shared similar annotations but it is through Denote’s [Acting_As] vocabulary that provided context information. For example, both correctly annotated “lettuce” to its surface form, but it was Denote that identified that lettuce was *acting as* a *main ingredient*. Similarly, Denote linked the phrase “*bacon, lettuce, and juicy tomato*” as an *alias* or *alternate name*.

TABLE II. ANNOTATION OF “BLT. THE […] IN EXISTENCE.” WITH DENOTE, TAGME AND DBPEDIA SPOTLIGHT.

Annotated Word(s)	Denote (DBPedia)	TagME (Wikipedia)	DBPedia Spotlight (DBPedia/Wikipedia)
BLT	Is_A {/BLT} Acting_As {/name}		/Bizarre_Love _Triangle
BLT sandwich	Is_A {/BLT} Acting_As {/name}	/BLT	
sandwich			/Sandwich
in existence			/Existence
sub		/Submarine_ sandwich	
crisp		/Potato_chip	
bacon	Is_A {/Bacon_sandwich, Bacon,Side_bacon} Acting_As {/mainIngredient, /ingredient}	/Bacon	
lettuce	Is_A {/Lettuce} Acting_As {/mainIngredient, /ingredient}	/Lettuce	
juicy		/Juice	
tomato	Is_A {/Tomato} Acting_As {/mainIngredient, /ingredient}	/Tomato	
bacon , lettuce and juicy tomato	Acting_As {/alias, /alternateName}		
scientific way		/Scientific_m ethod	

B. *The Number Annotator*

The number annotator is unique with respect to other annotators in that Denote attempts to identify text that is normally associated with a numerical value. Using statistical analysis on the Jaacard/Bayes-discovered list of similar concepts, Denote attempts to match up number values with annotated text. Figure 2 demonstrates on the input text “*The radio shack color computer has only 16 kb of memory*”.

“memory” With_Value 16 Of_Units #int Cat_Of
{/Home_Computers, TRS-80_Color_Computer}

Fig. 2. An example of number annotation with Denote

C. *The Categorizer*

Denote has access to over 656,000 categories defined in DBPedia’s <dcterm:subject> ontology. A Bayesian filter is used on each similar concept in order to determine if the subject(s) of which the concept belongs to is contextually related to the text being annotated. DBPedia Spotlight demo does not perform topic category determination. TagME’s demo performs topic categorization by simply listing its annotated text in a cloud-tag structure rather than a defined set of category topics. Consequently, we compare Denote’s output with Alchemy. The Alchemy annotator can perform named entity extraction from a list of 200+ defined (sub)-entities. In this comparison, the “storyline” of *The Godfather* movie was retrieved from the *Internet Movie Database* (IMDb) and annotated. Table III gives the results.

TABLE III. DENOTE VERSUS ALCHEMY IN CATEGORY/TOPIC TAGGING

Annotated Word(s)	Denote with Category Determination	Alchemy Entity Extraction
Corleone Family	Is_A {/The_Family_Corleone} Cat_Of {/Italian_American_novels, /Novels_about_organized_crime_in_the_United _States,/Novels_by_Mario_Puzo, /Family_saga_novels}	
Don		TelevisionShow
Vito Corleone	Is_A {/Vito_Corleone} Cat_Of {/The_Godfather_characters}	Person
Vito	Acting_As {Person}	
New York	Acting_As {Location}	City
Micheal		Person
Don Vito		Person
Don Vito Corleone	Is_A {/Don_Vito_Corleone} Cat_Of {/The_Godfather_characters}	
Don’s		Person
Mafia	Is_A {/Mafia_Don} Cat_Of {/The_Godfather_characters}	
Drugs	Is_A {/Drugs} Cat_Of {/The_Godfather_characters}	

Alchemy results were limited to primitive named entity types of city and person with the exception of an incorrect categorization of “television show”. In contrast, Denote tagged text into rich categories that include “Italian-American novels”, “organized crime novels”, and “Godfather characters ”.

IV. CONCLUSION

In this paper we demonstrated Denote – a semantic annotator based on the DBPedia ontology and compared its features with that of same-class text taggers. Denote’s middleware engine demo is available at http://ls3.rnet.ryerson.ca/annotator while a developer-friendly demo is at http://inextweb.com/denote_demo. Denote’s annotation capabilities are wrapped around a RESTful interface allowing for 3rd-party developers to create their own semantic-aware applications. The result, we hope, is an improvement in information search and retrieval for the end user. Our future work involves parallelisation to scale the service for a large number of concurrent clients. We are also developing proof-of-concept demonstrations including a semantic movie recommender whose database will be included as a data-set to the LOD project.

REFERENCES

[1] P. Ferragina, and U. Scaiella, “TAGME: On-the-fly Annotation of Short Text Fragments (by Wikipedia Entities)∗”, In Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM '10). 2010.

[2] P. Mendes, M. Jakob, A. García-Silva, and C. Bizer. “DBpedia spotlight: shedding light on the web of documents”, In Proceedings of the 7th International Conference on Semantic Systems (I-Semantics '11), 2011.

[3] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. “DBpedia - A crystallization point for the Web of Data.” In Web Semant. 7, 3 (September 2009), 154-165. 2009.