Extracting meaning from Dictionaries

Shreyas Shandilya and Krishna Raj

IIT Madras, Chennai TN 600036, India

Abstract. This work investigates the possible applications of dictionaries as a source of lexical knowledge. We have taken a graph theoretic approach to model the knowledge present in a dictionary as a Dictionary Network. To demonstrate possible applications of this network in NLP tasks we have devised an algorithms to retrieve documents related to a query. Using the hypothesis that meaning is a network property, the designed algorithm retrieves documents based on the relevance of the query to the document according to a meaning association measure that was formulated to take into account how meaning flows in the dictionary network. This paper also investigates the importance of words in a lexicon using pagerank algorithm.

Keywords: Dictionary Networks, Document Retrieval, Graphs, Depth First Traversal, Meaning Association

1 Introduction

A Dictionary is defined as a book or electronic resource that lists the words of a language (typically in alphabetical order) and gives their meaning, or gives the equivalent words in a different language, often also providing information about pronunciation, origin, and usage. The most prominent information that dictionaries contain are the word meaning relationships between words. A word is a linguistic representation of a concept in a language. All concepts in a language is the set of abstract objects that the language can be used to describe. Words are the representations that are used alone or together in a sentence to describe a concept or a number of concepts. Multiple words can be used to define a single concept (synonymy) and the same word can be used to define multiple concepts (polysymy). Words when combined together in phrases or sentences can be used to represent more complex concepts, formed by a combination of multiple.

Humans use a mental representation of the concept they know to generate language. A mental lexicon is used to find the most appropriate word/phrase that can represent the concept in question. The detailed process of how the manipulation and the retrieval of the appropriate words occur is not the topic of discussion of this paper. This idea of mental representations and lexicon can be extended to dictionaries, using concepts and words as their description. Concepts can be viewed as the mental representations that can be manipulated to create new concepts and the words in the dictionary form the lexicon to describe the concepts. This paper makes use of how a dictionary defines a word using other words and the idea of concepts and their representative words to investigate the use of dictionaries as a source of structured lexical information to perform NLP tasks. The application this paper focuses on is document retrieval using the meaning association between words.

1.1 Concepts and Words

Let C={c$_1$, c$_2$, c$_3$,.....,c$_n$} be the set of concepts described by the words in a dictionary D and W = {w$_1$, w$_2$, w$_3$,...., w$_n$} be the set of all words in D. The '*' operation combines two concepts to output a new more complex concept:

$$c_k = c_1 * c_2 \tag{1}$$

If words are being used to represent concepts in a language, a combination of words according to the aforementioned formulation will give a new concept:

$$c_k = w_1 * w_2 \tag{2}$$

$$w_1 * w_2 = c_1 * c_2 \tag{3}$$

where w$_1$ describes c$_1$ and w$_2$ describes c$_2$.

Thus, words manipulation of words can be performed to get the same concept as obtained by the combination of the concepts being represented by the two words. Thus, it can be concluded that,

$$w_k = w_1 * w_2 \tag{4}$$

Since concepts in a language are not explicitly available, manipulations can be performed on words to get new concepts. Dictionaries defined words using other words in a language. Each word in a language is used to described a concept. Thus, it can be said that '*' operation between two words will define a new word and this new word will then represent a concept formed by the combination of the concepts described by the words in the definition of the concerned word. This paper investigates the viability of application of this hypothesis in designing systems that can understand the meaning of words and understand meaning of sentences and phrases.

2 Prior Work

2.1 Dictionary Networks

The purpose of dictionaries, writes Wilks et al. (1993), "is to provide definitions of senses of words and, in so doing, supply knowledge about not just language, but the world." The definition of a word involves recursively new words, and thus, new senses and meanings. In this way, a dictionary can naturally be viewed as a network where each word w is related to the set of words w_1 , . . . , w_n that define it: for each word w so defined, consider the relationship w w_i for each i = 1, . . . , n, and proceed recursively with all the entries of the dictionary. These Dictionary Networks with proper modifications can be exploited for the purpose of word sense disambiguation, information retrieval, etc. In this paper we present a dictionary network extracted from Wiktionary. Dictionaries can be parsed to organize the knowledge with functions like Hypernym, Relation, Qualifier etc[**?**]. K.C. Litkowski (1978) was one of the first to state the importance of studying and exploiting dictionary networks, both as sources of material for natural language and to unravel the complexities of meaning. He presented three models to represent a dictionary. The first model uses nodes to represent words and edges to represent the relation w_a “is used to define” w_b . The second model extends the first one, adding relations between words and senses. In the third model, nodes represent concepts and edges represent different relations between them (senses, part of sentence, etc). Such models can be used a lexical resources, which can then be used to learn semantic relations, meaning association and language rules.

3 Methodology

3.1 Wiktionary

Wiktionary is a collaborative project licenced under both the Creative Commons Attribution-Share Alike 3.0 Un ported License and the GNU Free Documentation License. It is a free and open–source multilingual dictionary which now also includes a thesaurus, language statistics, etymologies, pronunciations, sample quotations, synonyms, antonyms and translations. In this project we used the English Wiktionary dump to extract the words and their definitions. The list of words is reduced by selecting only those words which are part of most frequent 1/3 million unigrams.

3.2 Network Construction

A graph G is defined by G = (V, E) where V is a set of n vertices and E is a set $E \subset V^2$ of m edges. V is the set of words and E is defined by a relation E $\rightarrow$ E : (w_1 , w_2) E if and only if w $\rightarrow$ w_2. An adjacency matrix A_m and an adjacency list A_l store the graph. Another vector W = {w_1,w_2,..., w_n} is the set of lemmatized words in the dictionary. Each edge e_i in the network will have a set of attributes Attr = $attr_1$, $attr_2$,....., $attr_k$. These attributes will be used to define the relationship R between two vertices.

The relationship being investigated in this paper is the meaning association that words have with the words that define them. As mentioned earlier, words define other words recursively, this relationship can be quantified using network properties and the properties of the paths between two words.

3.3 Meaning Association

A word is a single distinct meaningful element of speech or writing. Words or combinations of them can be considered as building blocks for more complex meaning structures. As mentioned in section 1.1, words can combine together to form more complex words and this new complex word will represent the concept that is defined by the combination of the words combining together. Levary et al (2012)[1] in their study of networks made from WordNet synsets, show that networks of graphs with meaning as the relationship between nodes show self-reference i.e. there are loops in the networks. These loops are introduced in the network when a new concept is introduced in the language. Due to the flow of meaning in such graphs, the loops are semantically coherent and words in loops share a similar meaning, i.e. they can be used to represent the same concept.

We build upon the work of Levary et al (2012) to hypothesize that loops in dictionary networks will also be semantically coherent. Unlike the graph made from synsets, the meaning within a loop will vary and as the size of loops become larger, the meaning association between two words decrease. We use this hypothesis to formulate a numerical value for meaning association between a word w and a word defining it, w$_{def}$. Let m be the meaning association between a w and w$_{def}$. Let L(w$_1$, w$_2$) be the shortest path from w$_1$ to w$_2$ such that

$$\begin{cases} L(w_1, w_2) = 1 & w_1 = w_2 \\ L(w_1, w_2) = 0 & \text{if no path} \\ L(w_1, w_2) = a & 0 \leq a \leq 1 \end{cases}$$

This requires correction. L(w1, w2) will be infinity in case of no path. 0 in case the words are the same and the length of the path in the case there exists a path between the two.

The formulation to quantify the value of meaning association between w and d$_{def}$ is as follows:

$$m = 1/(1 + L(w_{def}, w)) \tag{5}$$

This definition does not take into account the existence of loops

3.4 Query Expansion

Document retrieval using a query from a user makes use of a similarity between the query and the documents in a corpus to retrieve only the relevant documents. The most common method to compute the similarity between the query and the document is to compute the cosine similarity between a representation of the query and the document in the same vector space. Let q be a query and D be the set of all documents d$_1$, d$_2$,...,d$_n$. Let R be a method that convert the query q and document d$_k$ into a vector space such that,

$$s_k = R(q) * R(d_k) \tag{6}$$

where s$_k$ is the cosine similarity between words. Often there are documents in the corpus, which although do not have the same words in them as the query but have similar words and thus are relevant to the query. So a representation R must convert the query and the documents in the corpus into a vector space, such that the cosine similarity s$_k$ between R(q) and R(d$_k$) also takes into account all other possible words in the document, which are similar in meaning to the words in q, but are different.

Vector Representation To create an expanded vector representation for queries, we extract a sub-graph from the dictionary network, using depth first traversal in the dictionary network, which is terminated when the meaning association between a node and its neighbor goes below a threshold θ. This subgraph is then used to create a vector representation of the query. The vectors used to represent q and d$_k$ is of size $\|V\|whereVisthesetofallverticesinthedictionarynetwork.Therepresentationofd_k$ is a one hot encoded vector of all the words that are present in the document.

The vector representation R(q) is computed using the aforementioned extracted sub-graph. Since there are more than one ways a word can be connected to another word, if they are not neighbors. We define the following two formulations to find the meaning association between words:

$$m(w_k, w_l) = 1/(L(w_k, w_l) + L(w_k, w_l)) \tag{7}$$

Does not take into account the possibility of the forward and backward paths being the same. The same goes for the measure defined in eq 8. The concept of loops must be incorporated into the meaning association formulae.

where w$_k$ is the source node and w$_l$ is the node whose meaning association with the word w$_k$. w$_l$ here will be from the words originally in the query.

$$m(w_k, w_l) = \sum_{i \in P} \prod_{j=k}^{l} 1/(1 + L(w_{j+1}, w_j)) \tag{8}$$

where w$_k$ is the source node from the bag of lemmatized words of q and w$_l$ is the target node to compute the the meaning association with w$_k$. P is the set of all paths with disjoint nodes from w$_k$ to w$_l$

Equation 7 uses the same concept as the equation 5 to define the meaning association between two words. Equation 8 uses neighbour meaning association values to compute a global meaning association between two words. This formulation uses all possible ways in which w$_l$ can have its meaning associated with w$_k$ and gives an average of them all.

4 Future Work

References

1. Levary, David., Eckmann, Jean-Pierre., Moses, Elisha., Tlusty, Tsvi.: Loops and Self-Reference in the Construction of Dictionaries (2012)