Wikipedia-based Compact Hierarchical Semantics with Application to Semantic Relatedness

Sonya Liberman
Shaul Markovitch

Department of Computer Science
Technion - Israel Institute of Technology, 32000 Haifa, Israel

Abstract

A proper semantic representation of words and texts underlies many text processing tasks. In this paper, we present a novel representation of semantics which is based on an hierarchical ontology of natural concepts derived from Wikipedia articles and category system. Our method, called Compact Hierarchical Explicit Semantic Analysis (CHESA) generates compact hierarchical representations of unrestricted natural language texts. With comparison to previous methods for semantic representations, CHESA generates very intuitive and comprehensible representations allowing deep semantic reasoning and understanding. CHESA representations are flexible with regards to their level of abstraction and compactness. We present a methodology to compute semantic relatedness using CHESA representations and evaluate CHESA on the task of semantic relatedness assessment of words and texts. Empirical results show that for compact representations, CHESA is superior to the previous state of the art.

1. Introduction

Digital publicly available textual data is vast. The World Wide Web consists of tens of billions of web-pages, most of which contain textual content in the form of natural language. Intelligent access, manipulation and processing of such textual data is of great importance to a variety of tasks such as text categorization, word sense disambiguation and machine translation. One of the core issues of this challenge is representing language semantics in a way that can be manipulated by computers.

Statistical approaches for semantic representation of natural language perform a empirical analysis of textual data and construct statistical and probabilistic models of language. Most of these approaches do not utilize external textual data. They view a given text as a Bag of Words (BOW), namely, a text is represented as an unordered set of the words it contains. The BOW approach has been widely used in various information retrieval tasks such as search and text categorization (Baeza-Yates & Ribeiro-Neto, 1999; Sebastiani, 2002). Nevertheless, it performs sub-optimally on short texts, does little to address polysemy and synonymy and is not able to perform high level semantic reasoning.

It has long been recognized that in order to process textual data at a deeper level, computers require access to common-sense and domain-specific world knowledge (Buchanan & Feigenbaum, 1982; Lenat, Guha, Pittman, Pratt, & Shepherd, 1990). Indeed, considerable research efforts have been invested in developing methods for incorporating external knowledge into statistical NLP.
Corpus-based methods (Landauer & Dumais, 1997; Rehder, Schreiner, Wolfe, Laham, Landauer, & Kintsch, 1998; Landauer, Foltz, & Laham, 1998; Wolfe, Schreiner, Rehder, Laham, Foltz, Kintsch, & Landauer, 1998; Foltz, 1996; Burgess, Livesay, & Lund, 1998; Lund & Burgess, 1996; Turney, 2001) leverage large external textual corpora and use word-document co-occurrence information from these corpora. For example, the Latent Semantic Analysis (LSA) approach (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990) applies singular value decomposition to this co-occurrence data, retrieving “latent concepts” and representing words and texts in the space defined by these concepts. Various approaches (Budanitsky, Hirst, & Hirst, 2006; Jarmasz & Szpakowicz, 2003; Lee, Kim, & Lee, 1993; Jiang & Conrath, 1997; Leacock & Chodorow, 1998; Lin, 1998; Pedersen, Patwardhan, & Michelizzi, 2004; Wan & Peng, 2005) use lexicographic knowledge to represent semantics and assess semantic similarity. This knowledge is derived from lexical databases such as WordNet (Fellbaum, 1998) or Roget’s Thesaurus (Roget, 1852) encode important relations between words such as synonymy, hypernymy, and meronymy.

Other approaches utilize more extensive knowledge repositories to represent semantics and compute semantic relatedness. Some of these approaches (Strube & Ponzetto, 2006; Schonhofen, 2006; Milne, 2007; Kaiser, Schwarz, & Jakob, 2009) use Wikipedia, while others (Danushka, Yutaka, & Mitsuru, 2007; Matsuo, Sakaki, Uchiyama, & Ishizuka, 2006; Chen, Lin, & Wei, 2006; Sahami & Heilman, 2006; Maguitman, Menczer, Erdinc, Roinestad, & Vespignani, 2006; Lakkaraaju, Gauch, & Speretta, 2008; Gabrilovich & Markovitch, 2007a, 2005) utilize Web content and the Open Directory Project (ODP) repository.

The method of Explicit Semantic Analysis (ESA) introduced by Gabrilovich and Markovitch (Gabrilovich & Markovitch, 2006, 2007b, 2009) uses Wikipedia as a flat ontology, where each article represents an atomic concept. The semantics of a word is represented as a vector in the multi-dimensional space defined by Wikipedia articles. The weight of each dimension is inferred from the article text by computing the TF.IDF of the word within that text. A document is represented as the centroid of the representation vectors of its words. Experimental evaluation of ESA on semantic relatedness as well as text categorization tasks showed considerable improvements over previous approaches. Nevertheless, ESA has three notable drawbacks.

1. ESA representation is excessive. ESA represents word semantics as a weighted combination of all Wikipedia concepts that can amount to tens of thousands of concepts. Many text processing related tasks, e.g commercial search engines and real-time categorization systems need to process and index enormous amounts of textual content and then operate fast on these large corpora. The compactness of the semantic representations directly affects their performance, turning infeasible tasks to practicable and efficient.


Observe that eleven out of the top twenty concepts simply do not express the semantic meaning of the word money, as they refer to novels (The Color of Money), songs (I Get Money), television programs (The Money Programme) etc. As the association score for every concept is determined by the textual content of the corresponding article, the most highly associated concepts are roughly those that contain the word "money" many times.

2. ESA is noisy. Even in its compact form, ESA contains many redundant and over-specific concepts. ESA is not able to differentiate between concepts representing the primal gists of the word (such as Money for the word money in the example above), and over-specific, noisy concepts (such as Money, Money, Money that refers to a ABBA song), as the association scores for both types of concepts are high. As ESA generates each concept independently of the other concepts in the representation redundancies often occur. For example, the top-ten concepts generated by ESA for the word car, contain seven different car types (Concept car, Sports car, Armored car, Executive car, City car, Compact car and Full-size car) and two types of car number plates (Polish car number plates and Greek car number plates).

3. ESA if flat. ESA views Wikipedia as a flat conceptual ontology and constructs flat semantic representations, thus disregarding the inner structure and inter-dependencies of concepts. First, this is a notable weakness when addressing semantic relatedness of words and texts. When humans perform this task, they use their innate ability to generalize. For example, a human would easily determine that the words money and wealth are related as they both trigger high-level concepts related to economics and sociology. Observe the top twenty concepts generated by ESA for the word wealth:


Secondly, the ability to reason about semantics at varying abstraction levels is highly beneficial for many tasks. For example, given a text categorization task of assigning scientific papers to either Computer Science or History categories, the semantic representation of the texts can be captured on high-level. For instance, it is sufficient to determine whether the paper is in the field of applied sciences or humanities. Given a different task of categorizing papers to either supervised learning or unsupervised learning topics, the semantic representation of the texts should be much more specific.
Figure 1: The compact hierarchical semantic representation generated by CHESA for the word *money*

and fine-grained. Such multi-level, flexible reasoning cannot be achieved with a flat modeling of knowledge.

In this work we present a novel representation of semantics that overcomes these problems. Our method is called Compact Hierarchical Explicit Semantic Representation (CHESA). It represents semantics as a compact hierarchical structure of pre-defined natural concepts, capturing semantics at different abstraction levels and constructing representations at any given size, depending on the task at hand. Figure 1 shows CHESA semantic representation for the word *money*. The representation expresses the primal gist of the word, showing both specific (e.g., *Macroeconomics and monetary economics*) and high-level (e.g., *Society*) concepts triggered by it.

Numerous studies in cognitive sciences address the question of how knowledge is represented in memory. Many theoretical models assume that cognitive processes and representations have a hierarchical structure (Cohen, 2000). Behavioral (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976; Bower, Clark, Lesgold, & Winzenz, 1969; Potter & Faulconer, 1975) as well as neurological (Dennis, 1976; Hart, Berndt, & Caramazza, 1985) evidence support the notion of conceptual hierarchical structures in long-term memory. Studies on retention patterns of knowledge (Bahrick, 1984; Naveh-Benjamin, 1988; Conway, Cohen, & Stanhope, 1991; Cohen, Stanhope, & Conway, 1992; Stanhope, Cohen, & Conway, 1993) support the idea of inner hierarchical organization, as they reveal differences between the retention of higher level information and fine-grained specific details. Moreover, humans’ tendency to organize knowledge within hierarchies can be deduced from their handling of Web content. Yahoo! Directory\(^1\), the Open Directory Project\(^2\) (ODP), Wikipedia category system\(^3\) and numerous domain specific repositories are hierarchical ontologies, assembled

---

1. http://dir.yahoo.com
by humans in their attempt to organize information in a natural manner, easily perceived by them.

Our approach leverages structured encyclopedic knowledge encoded within Wikipedia articles and categories and uses the conceptual hierarchy inferred from this knowledge resource to represent text semantics. We start by defining a criterion for measuring the marginal contribution of a concept to the representation of a given word or text. We continue by presenting two greedy algorithms that construct compact hierarchical semantic representations of words based on that notion, and two extended algorithms that handle semantic representation of text. Finally, we define a similarity metric between hierarchical semantic representations, evaluate CHESA on the task of semantic relatedness assessment and show that it is superior to previous state of the art for compact representations.

The rest of this paper is organized as follows: Section 2 describes the methodology of CHESA and presents the algorithms for representing semantics of words and texts. Section 3 discusses the application of CHESA to automatic assessment of semantic relatedness. We evaluate CHESA and report on experimental results in Section 4. Related work is discussed in Section 5. We conclude in Section 6.

2. Compact Hierarchical Explicit Semantic Representation

Assume a pre-defined global hierarchical ontology, where nodes represent natural concepts and edges represent hyponymy or meronymy relations between these concepts. Assume that each node in that ontology is associated with some textual content. We represent the semantics of an input text \( t \) as a weighed sub-hierarchy within this global ontology. Namely, we draw a virtual separating curve on top of the hierarchy, that determines which concepts are included into the semantic representation (those above the separating curve) and which are excluded (those bellow the separating curve). Moreover, every concept within the representation is assigned a weight which represents the strength of association between its textual content and the input text.

In the rest of this section we address the two aspects that arise from our formulations. Namely, what knowledge repository can be used as a source for such global hierarchical ontology, and how to automatically construct a compact, comprehensive hierarchical semantic representation based on that ontology.

2.1 Wikipedia as a Conceptual Hierarchy

Wikipedia is the largest encyclopedia in the world. Its English version contains almost three million articles\(^4\) and a vast system of categories. It exceeds the size of the next largest English-language encyclopedia, Encyclopedia Britannica, by more than 25 times\(^5\). Wikipedia provides a comprehensive source of world knowledge, organized within a taxonomy-like structure determined by its articles and categories. Every article in Wikipedia contains a textual description of a single topic. We thus view each Wikipedia article as defining a concept corresponding to that topic. For example, the article titled *Dog* that describes the dog species, corresponds to the concept *Dog*. Articles in Wikipedia are usually classified

under one or more categories. Categories are also often classified under other categories. For example, the article *Cat* is classified under the category *Domesticated animals*, which is in turn, classified under the category *Animals*. We view Wikipedia categories as defining high-level concepts, which are more general than the concepts defined by the articles they categorize.

Wikipedia categories are typically established based on hyponymy and myronymy relations. Figure 2 shows a small subgraph of the Wikipedia category system. However, as a taxonomic category structure is not strictly enforced, cycles or disconnected components are possible (though rare). To simplify the original conceptual graph derived from Wikipedia, and eliminate cycles, we remove a small part of the edges in the graph. Specifically, if a node can be reached by more than one path from the root, we retain only the shortest path, using a breadth-first-search traversal through the graph. If several paths have the same minimal length, we retain all these paths. The specific algorithm can be found in Appendix ???. In the end of this procedure we obtain a rooted directed acyclic graph where inner nodes correspond to Wikipedia categories and leaf nodes correspond to Wikipedia articles.

We define the textual content of each concept \( v \) as follows:

\[
TC(v) = \begin{cases} 
\{A_v\} & \text{if } v \text{ is a leaf} \\
\bigcup_{v' \in \text{Children}(v)} TC(v') & \text{otherwise}
\end{cases}
\]

Where \( A_c \) is the article corresponding to the leaf concept \( c \) and \( \text{Children}(c) \) is the set of the direct child concepts of \( c \).
2.2 The Conditional Over-representation Criterion

Once the hierarchical ontology has been defined, the main question that needs to be addressed is which concepts from the ontology should be part of the semantic representation of a given text.

We begin by addressing the more simple case of constructing a semantic representation of a single word $w_i$. Potentially, every concept containing $w_i$ in one of the articles that constitute its textual content, can be part of the representation, as it is associated with the word to some extent. But once a certain concept is part of the representation, does it affect the marginal contribution of other concepts, specifically its direct descendants, to the representation? For instance, if the concept VEHICLES is part of the representation of the word car, how contributory is its child concept AUTOMOBILES to that representation? And what is the contribution of its child concepts SPACECRAFT or AMPHIBIOUS VEHICLES? Intuitively, the concept AUTOMOBILES provides a substantial value to the semantic interpretation of the word car even in the presence of the concept VEHICLES. The concept SPACECRAFT is unsuitable for describing the semantics of car. The concept AMPHIBIOUS VEHICLES is weakly associated with the word car, but it does not contribute to the grasping of the meaning of car beyond the interpretation provided by the concepts VEHICLES.

Based on those intuitions, we define the conditional over-representation criterion $\Phi_{w_i}(v_1, v_2)$ on an edge in the conceptual hierarchy. Its value quantifies the marginal contribution of the child concept $v_2$ given that its parent concept $v_1$ is included into the semantic representation. We wish to formalize this marginal contribution through a statistical test, and deduce upon its significance.

Recall that the textual content $TC(v_1)$, of some concept $v_1$, is a superset of the articles that constitute the textual content $TC(v_2)$, of its child concept $v_2$. We distinguish between two types of articles; those that contain the word $w_i$ in their text and those that do not. We then model $TC(v_2)$ as a sample from the population defined by $TC(v_1)$, with marked items being the articles containing the word $w_i$. Our null hypothesis postulates that $TC(v_2)$ is a random sample from the set of articles defined by $TC(v_1)$. We then use a hypergeometric test to compute the probability of falsely rejecting the null hypothesis. As this probability decreases, it is an indication that $w_i$ is over-represented in $v_2$ with respect to $v_1$.

A Hypergeometric test assumes a population of size $N$ in which $n$ items are marked, and a random sample of size $M$ drawn from that population. The number of marked items in the sample is described by a random variable $X \sim HG(N, n, M)$ that follows a hypergeometric distribution with a probability mass function defined as,

$$Pr(X = k) = \binom{n}{k} \frac{(N-n)\binom{M-k}{N-M}}{\binom{M}{N}}$$

The cumulative distribution function for the upper tail is

$$Pr(X \geq k) = \sum_{i=k}^{M} Pr(X = i)$$

Note that as $k$ increases, the probability of observing $k$ marked items or more in a random sample decreases.
Denote by $TC_w(v) \subseteq TC(v)$ the subsets of articles in the textual content of $v$ that contain the word $w$. Given an input word $w_i$ and a pair of concepts $(v_1, v_2)$ which constitutes an edge in the conceptual hierarchy we define the conditional over-representation criterion as

$$\Phi_{w_i}(v_1, v_2) = 1 - Pr(\mathcal{X}_{w_i,v_1,v_2} \geq |TC_{w_i}(v_2)|)$$

Given that

$$\mathcal{X}_{w_i,v_1,v_2} \sim HG(|TC(v_1)|, |TC_{w_i}(v_1)|, |TC(v_2)|)$$

$\Phi_{w_i}(v_1, v_2)$ is high when there is a low probability that a $|TC_{w_i}(v_2)|$ or more marked articles are found in a random sample. We can set a significance threshold on $\Phi_{w_i}(v_1, v_2)$ values, e.g., 0.95 to decide whether the concept $v_2$ should be included in the representation, given that $v_1$ is in the representation. We can also use $\Phi_{w_i}$ to rank concepts according to their marginal contribution to the representation.

For the aforementioned example of representing the semantics of the word car, the concept VEHICLES has 10,529 articles overall, out of which 2,232 contain the word car. The concept AUTOMOBILES has 1,936 articles, out of which 1,577 contain the word car. Even without performing a statistical test, it is evident that receiving this sample by chance is highly improbable, and indeed $\Phi_{\text{car}}(\text{Vehicles, Automobiles})$ is very close to 1.0 for these concepts. On the other hand, the concept SPACECRAFT contains the word car in only one out of its 705 articles. Consequently, the value of $\Phi_{\text{car}}(\text{Vehicles, Spacecraft})$ is very close to zero. The concept AMPHIBIOUS VEHICLES contains the word car in 8 out of 51 articles. In this case, the value of $\Phi_{\text{car}}$ is slightly higher ($\Phi_{\text{car}}(\text{Vehicles, Amphibious vehicles}) = 0.217$).

Note that the textual content of a leaf concept $v$ is based on exactly one article, $A_v$, namely, $|TC(v)| = 1$. Thus, we define the conditional over-representation criterion for a leaf concept and its parent concept as a function of the raw number of the input word occurrences within the textual content of the concepts$^6$. Let $\text{count}(w, v)$ indicate the number of times a word $w$ occurs in the textual content of a concept $v$. Then, given an input word $w_i$ and a pair of concepts $(v_1, v_2)$ where $v_2$ is a leaf, we define the conditional over-representation criterion as

$$\Phi_{w_i}(v_1, v_2) = 1 - Pr(\mathcal{X}_{w_i,v_1,v_2} \geq \text{count}(w_i, v_2))$$

Given that

$$\mathcal{X}_{w_i,v_1,v_2} \sim HG\left(\sum_j \text{count}(w_j, v_1), \text{count}(w_i, v_1), \sum_j \text{count}(w_j, v_2)\right)$$

2.3 Automatic Construction of Compact Hierarchical Semantic Representation

In this section we present two algorithms for constructing compact hierarchical representations of words. Both algorithms are based on the conditional over-representation criterion $\Phi$. Their modus operandi and consequently the semantic representations they generate, however, differ.

Both algorithms allow us to control the level of the compactness of the representation. To achieve compactness, semantics is described on a more abstract level, using general

---

$^6$ We also tried this approach for the internal nodes of the graph, but the results did not change significantly. Therefore, for the sake of efficiency, we decided to keep the original method.
concepts that appear high within the hierarchical ontology. As this constraint loosen, concepts located deeper within the hierarchy and expressing more specific topics, become part of the representation.

### 2.3.1 The Top-Down CHESA Algorithm

The Top-Down CHESA algorithm is an iterative greedy algorithm that traverses Wikipedia’s conceptual hierarchy in a top-down manner and expands the representation by adding more specific concepts located deeper in the hierarchy.

The process begins by representing the semantics of the input word $w_i$ by the root concept. Each iteration the algorithm investigates all the edges $(v_1, v_2)$ in the original hierarchy such that $v_1$ is in the representation, and $v_2$ has not yet been added. It then selects the edge $(v_1^{max}, v_2^{max})$ with the maximal value of $\Phi_{w_i}$. This edge, and the concept $v_2^{max}$ are then added to the representation.

The algorithm terminates when the representation reaches a pre-defined size of $k$ concepts. It can also receive a significance threshold for $\Phi_{w_i}$ as an input parameter. In that case, the algorithm terminates when all the edges considered in a given iteration have a $\Phi_{w_i}$ value below this threshold. The formal description of the procedure appears in Algorithm 1.

**Algorithm 1 Top-Down CHESA**

**Input:** An input word $w_i$, a rooted directed acyclic graph $G(V, E)$, a root node $r \in V$ and a representation size $k$

**Output:** A rooted directed acyclic subgraph $G'(V', E')$ of $k$ nodes

1. $V' \leftarrow \{r\}$
2. $E' \leftarrow \emptyset$
3. while $|V'| < k$ do
4. $(v_1^{max}, v_2^{max}) \leftarrow \arg \max_{(v_1,v_2)\in E,v_1\in V'} \Phi_{w_i}(v_1, v_2)$
5. $E' \leftarrow E' \cup \{(v_1^{max}, v_2^{max})\}$
6. $V' \leftarrow V' \cup \{v_2^{max}\}$
7. end while
8. return $G'(V', E')$

Note that the conceptual hierarchy may contain multiple inheritance. For this algorithm, it is sufficient for a concept to have a high $\Phi_{w_i}$ value with respect to at least one of its incoming edges to be included in the representation. Figure 2.3.1 shows the first five iterations of the top-down algorithm, for the word *money*, as well as its compact representation of size ten.

### 2.3.2 The Bottom-Up CHESA Algorithm

The way the top-down algorithm expands the representation, makes it susceptible to local minima. Specifically, when a concept has a low $\Phi_{w_i}$ value for all its out-going edges, its descendent concepts will not be added to the representation, even if the $\Phi_{w_i}$ values of *their*
out-going edges are high. One possible solution to this problem is to preform a lookahead. However, it is not clear how deep this lookahead should be.

Another solution, is reversing the representation construction process and operate in a bottom-up manner, by removing concepts rather than adding them. The Bottom-Up CHESA algorithm is an iterative greedy algorithm that handles this procedure. As opposed the Top-Down CHESA algorithm, it begins by representing the input word $w_i$ with all Wikipedia concepts, namely the full conceptual hierarchy. At each iteration, the algorithm investigates all the incoming edges of the leaf nodes in the current representation. It then selects the edge which is the least element under the ordering relation $\preceq_\Phi$ defined bellow, and removes it. If, as a consequence, the leaf node connected through that edge has zero in-coming edges in the representation, it is removed as well. The formal description of the procedure appears in Algorithm 2.

**Definition 1** Let $e = (v_1, v_2)$ and $e' = (v'_1, v'_2)$ be edges in the representation and $r$ be the root node. We define the ordering $\preceq_\Phi$ as: $e \preceq_\Phi e'$ if $\Phi_{w_i}(r, v_2) \geq 0.95$ and $\Phi_{w_i}(r, v'_2) < 0.95$ or $\Phi_{w_i}(v_1, v_2) \geq \Phi_{w_i}(v'_1, v'_2)$.

We do not simply select the edge with the minimal $\Phi_{w_i}$ since its value only indicates of the marginal contribution of the leaf node over its parent. Namely, $w_i$ cat be weakly
associated with the leaf node, but even more weakly associated with its parent, resulting in a high $\Phi_{w_i}$. We therefore, also consider the “absolute” contribution of the leaf by computing its conditional over-representation with respect to the root.

**Algorithm 2 Bottom-Up CHESA**

**Input:** An input word $w_i$, a rooted directed acyclic graph $G(V, E)$, a root node $r \in V$ and a representation size $k$

**Output:** A rooted directed acyclic subgraph $G'(V', E')$ of $k$ nodes

1: $V' \leftarrow V$
2: $E' \leftarrow E$
3: while $|V'| > k$ do
4:  Let $(v_{1\min}, v_{2\min})$ be the least element in $\{(v_1, v_2) \in E', \text{OutDegree}_{G'}(v_2) = 0\}$ under the ordering relation $\succeq_{\Phi}$
5:  $E' \leftarrow E' \setminus \{(v_{1\min}, v_{2\min})\}$
6:  if $\text{InDegree}_{G'}(v_{2\min}) = 0$ then
7:    $V' \leftarrow V' \setminus \{v_{2\min}\}$
8:  end if
9: end while
10: return $G'(V', E')$

It is evident that in many cases the top-down algorithm and the bottom-up algorithm produce different representations for the same input word and the same pre-defined representation size. In general, the top-down algorithm focuses on the primary meanings of the word and efficiently selects its most prominent gists within the conceptual hierarchy. However, the effect of local minima may lead to a situation where important concepts are absent from the representation. This may also be problematic when the topic coverage in Wikipedia is non-uniform. In that case, the top-down strategy sometimes focuses on only one, out of the several interpretations of the input word. The bottom-up algorithm aims to deal with these problems by trimming the representation at places where a generalization is suitable, but keeping concepts (and automatically their ancestors) which have a high $\Phi_{w_i}$ value with respect to their parent and the root concept.

Nevertheless, the bottom-up algorithm has some drawbacks with respect to the top-down technique. First, the complexity of generating a bottom-up representation is proportional to the size of the full conceptual hierarchy, while the complexity of the top-down strategy amounts to the pre-defined representation size. Moreover, all ancestors of a node that is not pruned from the representation, are present in the representation, regardless of their strength of association with the input word. Thus, the representations may contain many concepts which are weakly associated with the input word. The latter drawback of the bottom-up representation motivates an assignment of weights to the concepts, that quantify the strength of their association with the input word. We present our weighting scheme in the following section.

Figures 22 and 23, in Appendix B, show the top-down and the bottom-up representation of size 50 for the word *tiger*, respectively. Observe that the top-down representation is dominated by sports-related concepts, while the biological context of *tiger* is absent. This is due to the vast coverage of sports topics in Wikipedia, including numerous sport teams.
and clubs, many of which have the word tiger in their title. This non-uniform coverage, results in a low $\Phi_{tiger}(\text{MAIN TOPIC CLASSIFICATIONS,NATURAL SCIENCES})$ score and thus, concepts from this part of the hierarchy are not added to the representation. The bottom-up representation, on the other hand, contains the concepts NATURAL SCIENCES, BIOLOGY, ZOOLOGY, MAMMALOGY, MAMMALS and CARNIVORES, revealing the additional meaning of the word. Sports-related concepts are also present, but they are mostly high-level, and are not as abundant as in the top-down representation.

2.4 Assignment of Association Scores

Once a semantic representation is constructed, every concept in the representation is assigned a weight, indicating the strength of its association with the word being represented. This association score is a function of the frequency of the word in the textual content of the concept, and its global frequency in the conceptual hierarchy. We define the frequency of the word $w_i$ in the concept $v$ as

$$tf(w_i, v) = \frac{\text{count}(w_i, v)}{\sum_j \text{count}(w_j, v)}$$

We define the association score of an input word $w_i$ and a concept $v$ as follows:

$$s(w_i, v) = \max\left(0, \log \frac{tf(w_i, v)}{tf(w_i, r)}\right)$$

When $r$ is the root concept which contains the complete textual content of the hierarchy. Note that the association score is zero when the frequency of $w_i$ in the concept $c$ is equal to or lower than its frequency in the conceptual hierarchy, namely in the root.

2.5 Compact Hierarchical Semantic Representation of Texts

In the previous sections we described two algorithms for generating semantic representation of words. In the following, we present two extended algorithms for generating compact hierarchical semantic representations of texts.

2.5.1 The Union Algorithm

The Union CHESA algorithm represents an input text as the union of the compact representations of its words. Formally, given a text $t = \langle w_1, ..., w_n \rangle$, let $G_i(V_i, E_i)$ be a rooted directed acyclic graph, and $s(w_i, v), v \in V_i$ a weighting function, that constitute the CHESA representation of the word $w_i$. We define the representation of the text $t$ as a rooted weighted directed acyclic graph $G(V, E)$ such that:

$$V = \bigcup_{i=1}^{n} V_i$$

$$E = \bigcup_{i=1}^{n} E_i$$
We define the association score of the text $t$ and a concept $v$ as follows:

$$s(t, v) = \sum_{i=1}^{n} s(w_i, v)$$

Obviously, the union method can use both the top-down and the bottom-up algorithms for the intermediate step of representing single words.

### 2.5.2 The Direct Representation Algorithm

The Direct method represents the semantics of a text in a direct manner, by defining a new conditional over-representation criterion $\Phi$. Specifically, we modify the notion of marked items within a textual content of a concept, which previously referred to articles containing the target word to be represented.

Denote by $TC_{t,i}(v) \subseteq TC(v)$ the set of articles in the textual content of a concept $v$ that contain at least $i$ words from the input text $t$, when $i$ is a pre-defined parameter. We define the conditional over-representation criterion for a text $t$ and a pair of concepts $(v_1, v_2)$ as follows:

$$\Phi_t(v_1, v_2) = 1 - Pr(X_{t,v_1,v_2} \geq |TC_{t,i}(v_2)|)$$

Given that

$$X_{t,v_1,v_2} \sim HG(|TC(v_1)|, |TC_{t,i}(v_1)|, |TC(v_2)|)$$

Both the top-down and the bottom-up algorithms can be used to directly construct the semantic representation using the new definition of $\Phi$. Figure 4 shows the compact hierarchical representation of twenty concepts for the text fragment “A recently discovered type of star follows some wild orbits around the Milky Way”.

We define the association score of an input text $t = \langle w_1, ..., w_n \rangle$ and a concept $v$ as follows:

$$s(t, v) = \max \left( 0, \log \frac{\sum_{i=1}^{n} tf(w_i, v)}{\sum_{i=1}^{n} tf(w_i, r)} \right)$$

when $r$ is the root concept which contains the complete textual content of the hierarchy.

### 3. Using Compact Hierarchical Semantic Representation for Computing Semantic Relatedness

In this section we discuss the application of our novel hierarchical semantic representation methodology to automatic assessment of semantic relatedness.

Computing semantic relatedness between texts underlies many fundamental tasks in the field of text processing such as, text categorization, information retrieval and word sense disambiguation. In this section we present a two-phase, semantic approach for computing semantic relatedness. Given two words or texts, we generate their compact hierarchical semantic representations using the CHESA algorithms presented in Section 2. These representations are then compared, to deduce upon the semantic relatedness of the original texts.

---

8. The representation was constructed using the top-down algorithm and $i = 3$. 13
We define the semantic relatedness between two input texts, $t_1$ and $t_2$, as the cosine similarity between their linearized CHESA representations. Formally,

$$\text{rel}(t_1, t_2) = \frac{\sum_{v \in V} s(t_1, v)s(t_2, v)}{\sqrt{\sum_{v \in V} s^2(t_1, v)} \sqrt{\sum_{v \in V} s^2(t_2, v)}}$$

(1)

When $V$ is the set of all the concepts in the global hierarchy.

It seems that our definition of relatedness does not exploit the hierarchical structure of CHESA representations. Indeed, there exist other measures that use hierarchical information explicitly. Such measures, e.g. Generalized-cosine similarity (Ganesan, Garcia-Molina, & Widom, 2003) and Earth-mover’s distance (Rubner, Tomasi, & Guibas, 2000; Wan & Peng, 2005), compute similarity between collections, whose elements correspond to leaf nodes in some pre-defined global hierarchy. They then consider the relative proximity of the elements in that hierarchy to compute similarity. For example, the generalized-cosine similarity extends the cosine-similarity measure, by dropping the assumption that different elements correspond to orthogonal dimensions. It defines a positive inner product between different dimensions, which is a function of the proximity of the corresponding elements in the hierarchy.

We illustrate the contribution of explicitly considering hierarchical information in the following example. Assume a three level hierarchy as shown in Figure 5. Now, consider the following two vector pairs where the coordinates correspond to leaves $(a, b, c, d)$ in that hierarchy.

- $v_1 = (1, 0, 1, 0)$  $v_2 = (0, 1, 0, 1)$ (Figure 6(a))
- $u_1 = (1, 1, 0, 0)$  $u_2 = (0, 0, 1, 1)$ (Figure 6(b))

The cosine similarity for both vector pairs is zero. However, the general-cosine similarity, which considers the hierarchy, computes a positive similarity score for both pairs. Moreover,
it assigns a higher score to the first pair, as $a$ and $b$ (in $v_1$ and $v_2$ respectively) are located at greater proximity than $a$ and $c$ (in $u_1$ and $u_2$ respectively). The same occurs for $c$ and $d$ (in $v_1$ and $v_2$ respectively) compared to $b$ and $d$ (in $u_1$ and $u_2$ respectively).

(a) Pair 1 of collections

(b) Pair 2 of collections

Figure 6: Example of measuring similarity between collections which elements correspond to leaves of a hierarchy

Although it seems natural to use the aforementioned measures to compute similarity between CHESA representations, these measures are unsuitable for two reasons:

1. These measures assume a model where the collections they compare correspond to leaf nodes of some underlying hierarchical structure. CHESA representations do not coincide with such a model. Internal nodes of the Wikipedia conceptual hierarchy
are explicitly present within the representations. In fact, CHESA representations (in particular the compact ones) are mostly dominated by internal concepts and not the leaf nodes of the conceptual hierarchy.

2. CHESA algorithms ensure that a concept cannot be part of the representation unless its ancestors are part of it. By considering both internal and leaf nodes, cosine similarity is capable of implicitly capturing the hierarchical structure. As an example, observe the following two vector pairs where the coordinates correspond to all the nodes (A, B, C, a, b, c, d) in the hierarchy shown in Figure 5.

- \( v_1' = (1, 1, 1, 1, 0, 1, 0) \)
- \( v_2' = (1, 1, 1, 0, 1, 0, 1) \) (Figure 7(a))
- \( u_1' = (1, 1, 0, 1, 1, 0, 1) \)
- \( u_2' = (1, 0, 1, 0, 0, 1, 1) \) (Figure 7(b))

Note that the elements defined by \( v_1', v_2', u_1' \) and \( u_2' \) are a combination of the elements defined by \( v_1, v_2, u_1 \) and \( u_2 \) respectively, and their appropriate ancestors. The cosine similarity for both pairs is positive, and it is higher for the first pair, which coincides with the results achieved by the generalized cosine similarity measure in the previous example.

![Diagram](image.png)

Figure 7: Example of measuring similarity between collections which elements correspond to leaves of a hierarchy and their ancestor inner nodes

Using a cosine similarity for assessing semantic relatedness is also advantageous in terms of computational time complexity, as it is linear in the size of the representations, while using measures that exploit proximities within a hierarchy amounts to at least a quadratic complexity. In terms of space complexity, there is no need to store the complete Wikipedia
conceptual hierarchy, but only the CHESA representations of the texts, meaning that compact representations amounts to fast computation of semantic relatedness and low storage costs.

4. Empirical Evaluation

In this section we evaluate CHESA by performing both quantitative and qualitative analysis of the method on the task of semantic relatedness assessment, and compare it to ESA, the previous state of the art, as well as other previous work. Note that semantic relatedness and semantic similarity are different notions. It is possible that two terms are related but not similar, e.g., Obama and Washington are not similar but are considered to be related. We address the issue of semantic similarity in Section 4.6.

4.1 Experimentation Procedure

Assessing semantic relatedness of words and texts is a difficult task for computers but a natural, everyday tasks for humans. Moreover, studies have shown that the inter-agreement between different individuals on such tasks is very high (Budanitsky et al., 2006; Jarmasz & Szpakowicz, 2003; Finkelstein, Gabrilovich, Matias, Rivlin, Solan, Wolfman, & Ruppin, 2002). This is not surprising as it is this agreement that allows us to understand each other and communicate. Therefore, we consider human judgements as a golden standard and evaluate our method in terms of the correlation of the computed relatedness scores with the scores assigned by humans.

We use two benchmarks for assessing semantic relatedness of words and texts. The WordSimilarity353 collection (Finkelstein, Gabrilovich, Matias, Rivlin, Solan, Wolfman, & Ruppin, 2001; Finkelstein et al., 2002) is assembled of 353 noun pairs. Each pair is judged by 13-16 judges, which are university graduates with fluent knowledge of the English language. The relatedness scores are given in a scale of 0 (completely unrelated words) to 10 (very much related or identical words), and they are averaged for each word pair to produce a single score.

The Lee collection (Lee, Pincombe, & Welsh, 2005; Pincombe & Surveillance, 2004) consists of 50 documents from the Australian Broadcasting Corporation’s news mail service. The documents are paired in all possible ways and judged for semantic relatedness by 83 university students. The judgements are on a five-point scale with 1 indicating highly unrelated texts and 5 indicating highly related texts.

To compute correlation with human judgements in the WordSim353 benchmark, we use the Spearman’s rank-order correlation coefficient. Being non-parametric, it is considered to be more robust than Pearson’s linear correlation coefficient. Note, however, that the calculations in Spearman’s measure become inaccurate as the number of tied ranks increases. Thus, Spearman’s correlation is not appropriate for the Lee benchmark, where there are only 67 distinct values among the 1,225 relatedness scores. Therefore, we use Pearson’s linear correlation to evaluate CHESA on this benchmark.

The aforementioned collections are, to the best of our knowledge, the largest publicly available collections for assessment of semantic relatedness, and are widely used to evaluate methods for assessing semantic relatedness (Jarmasz & Szpakowicz, 2003; Finkelstein et al., 2002; Strube & Ponzetto, 2006; Hughes & Ramage, 2007; Gabrilovich & Markovitch, 2007b;
Liberman and Markovitch

<table>
<thead>
<tr>
<th>Number of nodes</th>
<th>557,114</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of inner nodes</td>
<td>79,886</td>
</tr>
<tr>
<td>Number of leaf nodes</td>
<td>477,228</td>
</tr>
<tr>
<td>Number of distinct terms</td>
<td>730,489</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Average</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Out-degree</td>
<td>17.16</td>
<td>1</td>
<td>5,041</td>
</tr>
<tr>
<td>In-degree</td>
<td>2.46</td>
<td>1</td>
<td>51</td>
</tr>
<tr>
<td>Depth</td>
<td>6.59</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>Number of distinct terms per article</td>
<td>265.85</td>
<td>100</td>
<td>4,216</td>
</tr>
<tr>
<td>Number of terms per article</td>
<td>668.12</td>
<td>115</td>
<td>462,076</td>
</tr>
</tbody>
</table>

Table 1: Properties of Wikipedia’s conceptual hierarchy based on October 18, 2007 snapshot

Lee et al., 2005). Additional benchmarks exist, which assess semantic similarity rather than semantic relatedness, e.g., the Li sentence collection (Li, McLean, Bandar, O’Shea, & Crockett, 2006). These benchmarks will be discussed in Section 4.6.

We implemented our CHESA approach using a Wikipedia snapshot as of October 18, 2007. Wikipedia is publicly available for download in an XML format. We parsed its XML dump using the Wikipedia Processor (WikiPrep) tool (Gabrilovich & Markovitch, 2006, 2007b) and obtained 6.8 Gigabyte of text. Our data includes over 2 million Wikipedia articles and almost 300,000 categories.

We follow the footsteps of Gabrilovich and Markovitch (Gabrilovich & Markovitch, 2006) and discard overly-small articles having fewer than 100 non stop words, and articles having less than 5 incoming or outgoing links. We also exclude other types of articles and categories that are unlikely to be useful for describing semantics. For example articles and categories referring to a specific date or an event occurring in a specific year (e.g. 1961 PLAYS, 1820S IN FASHION) or categories which clearly do not express hypernymy relations (e.g. WRITERS BY NATIONALITY and BRITISH MUSICIANS BY INSTRUMENT). We also disregard lists and stubs. At the end of this process our knowledge base contained 497,153 articles and 125,542 categories. We then used the procedure described in Section 2.1 to construct a conceptual hierarchy rooted at the MAIN TOPIC CLASSIFICATIONS concept.

Table 1 summarizes the properties of the resulting conceptual hierarchy. To compute the statistics on the out-degree we considered only the internal nodes. To compute the statistics on node depth, we considered only the leaf nodes.

4.2 Evaluating CHESA on Semantic Relatedness of Words

In this section we report the results of an experimental evaluation of CHESA on semantic relatedness of words. We evaluate the performance of CHESA on the WordSimilarity353 collection and perform two experiments. In the first experiment, we use the top-down CHESA algorithm (Section 2.3.1) to generate compact representations, given different, pre-defined values of representation size $k$. In the second experiment we use the bottom-up

CHESA algorithm (Section 2.3.2) to generate the representations. For both experiments we use Equation 1 to compute semantic relatedness.

For comparison, we compute semantic relatedness using ESA compact representations by taking the top-$k$ most highly-associated articles from the ESA interpretation vector\textsuperscript{10}. Figure 8 shows the performance of the algorithms with varying values of representation size.

We see that both the top-down and the bottom-up CHESA algorithms significantly outperform ESA for compact representations\textsuperscript{11}. We believe that the advantage of CHESA is the hierarchical structure of its representations, which captures semantics on different abstraction levels, thus better revealing the various relations between the input words. In Section 4.5 we perform a qualitative analysis that demonstrates this phenomena.

The results also show that with tighter constraints on representation size, the top-down CHESA algorithm is superior to the bottom-up algorithm. As representation size increases, the bottom-up algorithm outperforms the top-down algorithm although the difference is not statistically significant.

In addition to experimenting with varying values of $k$, we evaluate CHESA algorithms in a setting of unlimited representation size, namely we represent semantics with the full hierarchical representation. Note that the top-down and the bottom-up CHESA algorithms

\textsuperscript{10} Both CHESA and ESA representations are constructed based on the same Wikipedia snapshot from October 18, 2007.

\textsuperscript{11} The difference is statistically significant with $p$-value $< 0.05$ for $k \leq 100$ for top-down CHESA and $k \leq 200$ for bottom-up CHESA.
<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>WordNet (Jarmasz &amp; Szpakowicz, 2003)</td>
<td>0.35</td>
</tr>
<tr>
<td>Rogets Thesaurus (Jarmasz &amp; Szpakowicz, 2003)</td>
<td>0.55</td>
</tr>
<tr>
<td>LSA (Finkelstein et al., 2002)</td>
<td>0.56</td>
</tr>
<tr>
<td>WikiRelate! (Strube &amp; Ponzetto, 2006)</td>
<td>0.50</td>
</tr>
<tr>
<td>MarkovLink (Hughes &amp; Ramage, 2007)</td>
<td>0.55</td>
</tr>
<tr>
<td>CHESA</td>
<td>0.72</td>
</tr>
<tr>
<td>ESA (Gabrilovich &amp; Markovitch, 2007b)</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Table 2: Computing semantic relatedness of words when representation size is unlimited.

produce identical semantic representations of words when representation size is unlimited. For comparison, we compute semantic relatedness using ESA without trimming its interpretations vectors.

Table 2 shows the results of both CHESA algorithms, ESA and five additional previous methods\(^\text{12}\). We see that when representation size is unlimited, ESA slightly outperforms both CHESA algorithms, however, the difference is not statistically significant. ESA and CHESA yield significant improvements\(^\text{13}\) over the other methods. Note that CHESA compact representations of size above 100 concepts outperform these previous methods.

4.3 Evaluating CHESA on Semantic Relatedness of Texts

To evaluate CHESA on the assessment of semantic relatedness of texts, we use the Lee document collection and conduct four experiments. In the first two, we generate compact representations of the texts using the Union algorithm (Section 2.5.1) based on the top-down and the bottom-up representations of words. In the other two experiments we generate CHESA representations using the Direct algorithm (Section 2.5.2), again, using both the top-down and the bottom-down techniques. For all experiments we use Equation 1 to compute semantic relatedness.

Recall that the Union algorithm generates a representation for an input text by first representing each word of the text with a compact representation of size $k$, and then unifying these representations. The size of the final representation of the text is thus between $k$ and $nk$, when $n$ is the number of words in the text. The Direct algorithm, on the other hand, constructs the semantic representation directly, without an intermediate step of representing the single words of the text. It uses a modified $\Phi$ function, and applies either the top-down or the bottom-up algorithms to construct a representation of a pre-defined size $k$.

For comparison, we also compute semantic relatedness using ESA representations of the texts. Note, that ESA represents texts as a centroid of the interpretation vectors of its words. It first represents the semantics of each word within the text, and then averages the vectors generated for each word. To generate compact representations using ESA, we can either prune the interpretation vectors to a pre-defined size $k$ at the level of single words, or, at the level of the final text interpretation vector (using the full vectors when representing

\(^{12}\) The results of the five additional previous methods are taken from (Gabrilovich & Markovitch, 2007b)

\(^{13}\) The difference is statistically significant with p-value $< 0.05$
The Union algorithm is more similar to the former version of ESA, and thus we use this version when comparing ESA to Union CHESA. The Direct algorithm is more similar to the latter version, and therefore compared to it.

Figure 9 shows the result for evaluating Union CHESA and ESA algorithms. Note that when comparing the Union algorithms to ESA, we use an inner parameter $k$ that specifies the size of the representations of each word in the text. The final, text representations for both Union CHESA and ESA are of size between $k$ and $nk$ (where $n$ is the number of words in the text). The results show the performance of the algorithms for $k = 10, 50, 100, 200, 300, 500, 1000$, where for each $k$, we compute the average representation size for all the documents in the Lee collection, as well as the correlation score with human judgements.

We see that in general, both Union CHESA algorithms outperform ESA and that the top-down algorithm is superior to the bottom-up algorithm. We note that for very small representation size ESA outperforms CHESA. We hypothesize that this phenomenon occurs since CHESA achieves compactness by using more abstract concepts. While this very compact description is often sufficient to express the semantics of single words (CHESA outperforms ESA for very compact representations in the task of word relatedness), it is not able to efficiently capture the semantics of a complex text.

Figure 10 shows the results for evaluating the Direct CHESA and the ESA algorithms\textsuperscript{14}. We see that the bottom-up algorithm outperforms ESA for compact representations of 50

\textsuperscript{14} For the Direct CHESA algorithm we set $i = 5$
Figure 10: Using Direct-CHESA for semantic relatedness of texts with varying representation size

Table 3: Computing semantic relatedness of texts when representation size is unlimited

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bag of words (Lee et al., 2005)</td>
<td>0.35</td>
</tr>
<tr>
<td>LSA (Lee et al., 2005)</td>
<td>0.55</td>
</tr>
<tr>
<td>Union CHESA</td>
<td>0.60</td>
</tr>
<tr>
<td>Direct CHESA</td>
<td>0.70</td>
</tr>
<tr>
<td>ESA (Gabrilovich &amp; Markovitch, 2007b)</td>
<td>0.72</td>
</tr>
</tbody>
</table>

concepts\textsuperscript{15}, and the gap between them diminishes as representation size increases. The performance of the top-down algorithm is comparable to ESA for representations above the size of 50 concepts. For very compact representations of size 10, ESA outperforms both Direct CHESA algorithms.

In addition, we evaluate the Union and the Direct CHESA algorithms in a setting of unlimited representation size. The results are shown in Table 3. We see that ESA slightly outperforms the Direct CHESA algorithm, however, the difference is not statistically significant. Both algorithms outperform the Union CHESA algorithm and previous methods (p-value < 0.05).

\textsuperscript{15} The difference is significant at p-value < 0.05 for $k = 50$
4.4 The Effect of Different Distance Measures on CHESA Performance

In Section 3 we discussed the various possibilities to compute the distance between two CHESA representations and claimed that the cosine similarity measure captures the hierarchical inner structure of the representations. To test our hypothesis we compared cosine similarity to generalized cosine similarity (Ganesan et al., 2003). The generalized cosine similarity measure takes into account only the leaf concepts of the representations, and computes a similarity score based on these concepts and their proximities in the global hierarchy. We also tested the importance of using the association scores of concepts, for computation, by comparing cosine similarity to Jaccard measure (Jaccard, 1901) that disregards weighting.

To assess semantic relatedness using the cosine similarity and Jaccard measures we consider both the internal and the leaf concepts within the CHESA representations. When using general cosine similarity, we consider only the leaf concepts (these are the concepts for which none of the children is in the representation).

Figures 11(a) and 11(b) show the results for assessing relatedness of words with the three similarity measures, using the top-down and bottom-up algorithms respectively. Figures 11(c) and 11(d) show the results for assessing relatedness of texts with the three similarity measures, using the top-down and bottom-up algorithms respectively.

It is seen from the graphs that the cosine similarity and Jaccard measures are superior to the generalized cosine similarity measure for both words and texts. Cosine similarity significantly outperforms Jaccard on assessing semantic relatedness of words with the bottom-up algorithm, and is slightly superior to it in the other three experiments. These results support our initial choice of cosine similarity as a measure of comparing CHESA representations. They stress that applying cosine similarity to hierarchical representations is advantageous over explicitly using the underlying hierarchical structure, but considering the leaf nodes only. Moreover, it illustrates the importance of assigning concept association scores.

4.5 Qualitative Analysis of CHESA

To better understand the specific properties of CHESA and illustrate its strength in generating compact comprehensive representations and assessing semantic relatedness, we present a qualitative analysis of our method using real examples.

1. **Money vs. Wealth**

   Consider the words *money* and *wealth* from the WordSimilarity353 collection. This pair was rated by human judges as highly related (among the top 40 related pairs). Figure 1 shows the CHESA top-down representations of size 10 for these words.

   We see that in a limited space of ten concepts, CHESA representation captures the core semantic gists of the words on several abstraction levels. The two representations share several concepts (marked in bold) such as Economics and Sociology indicating of the semantic relatedness between the two words.

   For comparison, Table 4 shows the top-10 concepts generated by ESA for each of the words. We can see that the ESA interpretation vectors for these two words do not
Figure 11: The affect of different similarity measures on CHESA performance for semantic relatedness

share any concepts and thus the relatedness score computed using cosine similarity is zero.

2. Vodka vs. Brandy

Another pair that is judged by humans as very related in the WordSimilarity353 collection is vodka and brandy. Figure 2 shows the top-down representations of size 10 of the two words. The two representation share many concepts, such as ALCOHOLIC BEVERAGES, MIXED DRINKS and their ancestors. The concepts VODKA and BRANDY are unique to each representation and indicate of the differences in the semantics of the words.

Table 5 shows the top-10 concepts generated by ESA for the two words. We see that the generated concepts are over-specific and that the compact representations have no intersecting concepts. Thus, ESA in unable to capture the relations between the words. Moreover, the representation of the word brandy has only 1 concept associated with its main context of an alcoholic beverage, while the other concepts refer to people.
<table>
<thead>
<tr>
<th>Money</th>
<th>Wealth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Money</td>
<td>Wealth in the United States</td>
</tr>
<tr>
<td>Money, Money, Money</td>
<td>Wealth</td>
</tr>
<tr>
<td>Money creation</td>
<td>Distribution of wealth</td>
</tr>
<tr>
<td>Money for Nothing/Beverly Hillbillies</td>
<td>Wealth management</td>
</tr>
<tr>
<td>Make Money Fast</td>
<td>Wealth (economics)</td>
</tr>
<tr>
<td>Money supply</td>
<td>Sovereign wealth fund</td>
</tr>
<tr>
<td>Moneyness</td>
<td>The Wealth of Nations</td>
</tr>
<tr>
<td>Cash Money Records</td>
<td>Land of Wealth</td>
</tr>
<tr>
<td>Eddie Money</td>
<td>Common Wealth Party</td>
</tr>
<tr>
<td>Money laundering</td>
<td>Share Our Wealth</td>
</tr>
</tbody>
</table>

Table 4: ESA top-10 generated concepts for the words *money* and *wealth*

<table>
<thead>
<tr>
<th>Vodka</th>
<th>Brandy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vodka</td>
<td>Brandy</td>
</tr>
<tr>
<td>Absolut Vodka</td>
<td>Albrecht Brandi</td>
</tr>
<tr>
<td>Smirnoff</td>
<td>George Brandis</td>
</tr>
<tr>
<td>SKYY vodka</td>
<td>Kristina Brandi</td>
</tr>
<tr>
<td>Bloody Mary (cocktail)</td>
<td>Brandy (entertainer)</td>
</tr>
<tr>
<td>Gin (Case Closed)</td>
<td>Brandi Chastain</td>
</tr>
<tr>
<td>Mixed drink shooters and drink shots</td>
<td>Jonathan Brandis</td>
</tr>
<tr>
<td>Vodka Belt</td>
<td>Brandy &amp; Mr. Whiskers</td>
</tr>
<tr>
<td>Grey Goose (vodka)</td>
<td>Brandi Carlile</td>
</tr>
<tr>
<td>Beer cocktail</td>
<td>Tom Brandi</td>
</tr>
</tbody>
</table>

Table 5: ESA top-10 generated concepts for the words *vodka* and *brandy*
3. Synonyms
The former examples showed cases of words which are semantically related. To see how CHESA handles synonyms we present the representations of the words car and automobile in Figure 3.

We see that both representation share the concept VEHICLES and its ancestors, which are indeed the connecting link between the two words. However, while car and automobile are considered synonyms\(^\text{16}\), their CHESA representations differ. For the word car, the concepts related to sports are prominent, while the word automobile is represented exclusively with concepts related to technology. These differences reflect the differences in the language usage of the two words, as the word automobile is used less in the context of sports and car racing.

Table 6 shows the top-10 concepts generated by ESA for both words. We see that ESA generates overly-specific concepts for the two words. For example, among the top-10 concepts generated for car, 7 refer to different car types and 2 refer to two kinds of car number plates. It is thus unable to capture the relations between them.

\(^{16}\) http://thesaurus.reference.com/browse/car
Figure 13: Comparison of the compact top-down CHESA representations of the words \textit{vodka} and \textit{brandy}.

4. \textbf{Smile vs. Chord}

The aforementioned examples stress the importance of generalization when representations are compact. As ESA is unable to generalize, salient semantic relations are often missed. Nevertheless, in some cases, generalization may lead to unexpected results. Figure 4 shows the top-down representations of size 10 for the words \textit{smile} and \textit{chord}. This pair of words was taken from the WordSimilarity353 collection, where it was judged by humans as being very unrelated.

CHESA, on the other hand, finds the two words to be very related, and consequently fails when compared to human judgements. We see that although the representation of \textit{smile} contains the concepts \textit{Society} and \textit{Communication}, which are related to the main gists of the word, it also contains many music-related concepts. The reason for that is the vast amount of music-related Wikipedia articles, such as songs and bands, that contain the word.

5. \textbf{Idioms}

The word pair (\textit{death,row}) appears in the WordSimilarity353 collection and is judged by humans as very related. The reason is most probably the fact that this pair is
(a) Compact representation of car

(b) Compact representation of automobile

Figure 14: Comparison of the compact top-down CHESA representations of the words car and automobile

recognized as the idiom “death row”. Figure 5 shows the top-down representations of size 10 for each of the words. We see that CHESA does not identify these two words as related. The semantic interpretation of the word row contains concepts such as Linear Algebra and Principal Component Analysis, while the representation of the word death contains concepts such as Life, Death and Men. Indeed, an idiom is defined as “an expression whose meaning is not predictable from the usual meanings of its constituent elements”\textsuperscript{17}. CHESA’s assessment of the semantic relatedness is correct in this example, although it does not coincide with human judgements, as

\textsuperscript{17} http://dictionary.reference.com/browse/idiom
Table 6: ESA top-10 generated concepts for the words *car* and *automobile*

<table>
<thead>
<tr>
<th>Car</th>
<th>Automobile</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Polish car number plates</td>
<td>Automotive industry</td>
</tr>
<tr>
<td>2 Concept car</td>
<td>Automobile</td>
</tr>
<tr>
<td>3 Sports car</td>
<td>Federation Internationale de l’Automobile</td>
</tr>
<tr>
<td>4 Armored car</td>
<td>Automotive design</td>
</tr>
<tr>
<td>5 Greek car number plates</td>
<td>American Automobile Association</td>
</tr>
<tr>
<td>6 Executive car</td>
<td>California Automobile Association Building</td>
</tr>
<tr>
<td>7 City car</td>
<td>Monte Carlo Rally</td>
</tr>
<tr>
<td>8 Compact car</td>
<td>Automobile Club de l’Ouest</td>
</tr>
<tr>
<td>9 Full-size car</td>
<td>Car accident</td>
</tr>
<tr>
<td>10 Car bomb</td>
<td>Automobile Magazine</td>
</tr>
</tbody>
</table>

the relation between the two words is not semantic but rather associative. ESA, on the other hand, often identifies associative relations, since words which are part of an idiom co-occur very often. For example, the ESA top-10 representations of the words *death* and *row* both contain the concept *Death row*.

6. **Semantic relatedness of texts**

Figure 17 shows the CHESA compact representation of size 20 for the texts “A recently discovered type of star follows some wild orbits around the Milky Way” and “Astronomers measured the distance to a faraway galaxy, to investigate universe expansion”.

First we see that both representations capture the astronomy-related primal gists of the text fragments, at different levels of abstraction. We see that some concepts, such as, *Astronomy*, *Universe* and *Space* are shared between the representations. These concepts express well the relatedness between the two text fragments. Other concepts are unique to each representation, and stress the semantic differences between the two texts. For example, the concept *Extrasolar planet* appears only in the representation of the first text, while the concept *Relativity* appears only in the representations of the second text.

Table 7 shows the top-20 concepts generated by ESA for each text fragment. Although these representations share several concepts such as *Milky Way*, *Supernova* and *Andromeda Galaxy*, the intersecting concepts are overly-specific and weakly express the relation between the two text fragments.

7. **Word Sense Disambiguation**

Figure 18 shows the compact top-down representation of size 20 of the ambiguous word *bank*. We see that the representation contains concepts expressing both meanings of the word (e.g., the concepts *Financial economics* for the financial context and *Landforms* for the geographical context). Nevertheless, the financial context is more

---

18. Representation is generated using the top-down Direct CHESA with \( i = 3 \)
“A recently discovered type of star follows some wild orbits around the Milky Way”

“Astronomers measured the distance to a faraway galaxy, to investigate universe expansion”

<table>
<thead>
<tr>
<th></th>
<th>“Astronomers measured the distance to a faraway galaxy, to investigate universe expansion”</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Milky Way</td>
</tr>
<tr>
<td>2</td>
<td>Andromeda Galaxy</td>
</tr>
<tr>
<td>3</td>
<td>Globular cluster</td>
</tr>
<tr>
<td>4</td>
<td>Star cluster</td>
</tr>
<tr>
<td>5</td>
<td>Sagittarius Dwarf Elliptical Galaxy</td>
</tr>
<tr>
<td>6</td>
<td>Binary star</td>
</tr>
<tr>
<td>7</td>
<td>Extrasolar planet</td>
</tr>
<tr>
<td>8</td>
<td>Solar System (1945 play)</td>
</tr>
<tr>
<td>9</td>
<td>Canis Major Dwarf Galaxy</td>
</tr>
<tr>
<td>10</td>
<td>Astrometry</td>
</tr>
<tr>
<td>11</td>
<td>Spiral galaxy</td>
</tr>
<tr>
<td>12</td>
<td>Red dwarf</td>
</tr>
<tr>
<td>13</td>
<td>Supernova</td>
</tr>
<tr>
<td>14</td>
<td>Comet</td>
</tr>
<tr>
<td>15</td>
<td>Variable star</td>
</tr>
<tr>
<td>16</td>
<td>Meteoroid</td>
</tr>
<tr>
<td>17</td>
<td>Parsec</td>
</tr>
<tr>
<td>18</td>
<td>Black hole (1945 play)</td>
</tr>
<tr>
<td>19</td>
<td>Space science</td>
</tr>
<tr>
<td>20</td>
<td>Dial-Home Device</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>“A recently discovered type of star follows some wild orbits around the Milky Way”</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Milky Way</td>
</tr>
<tr>
<td>2</td>
<td>Andromeda Galaxy</td>
</tr>
<tr>
<td>3</td>
<td>Globular cluster</td>
</tr>
<tr>
<td>4</td>
<td>Star cluster</td>
</tr>
<tr>
<td>5</td>
<td>Sagittarius Dwarf Elliptical Galaxy</td>
</tr>
<tr>
<td>6</td>
<td>Binary star</td>
</tr>
<tr>
<td>7</td>
<td>Extrasolar planet</td>
</tr>
<tr>
<td>8</td>
<td>Solar System (1945 play)</td>
</tr>
<tr>
<td>9</td>
<td>Canis Major Dwarf Galaxy</td>
</tr>
<tr>
<td>10</td>
<td>Astrometry</td>
</tr>
<tr>
<td>11</td>
<td>Spiral galaxy</td>
</tr>
<tr>
<td>12</td>
<td>Red dwarf</td>
</tr>
<tr>
<td>13</td>
<td>Supernova</td>
</tr>
<tr>
<td>14</td>
<td>Comet</td>
</tr>
<tr>
<td>15</td>
<td>Variable star</td>
</tr>
<tr>
<td>16</td>
<td>Meteoroid</td>
</tr>
<tr>
<td>17</td>
<td>Parsec</td>
</tr>
<tr>
<td>18</td>
<td>Black hole (1945 play)</td>
</tr>
<tr>
<td>19</td>
<td>Space science</td>
</tr>
<tr>
<td>20</td>
<td>Dial-Home Device</td>
</tr>
</tbody>
</table>

Table 7: ESA top-20 generated concepts for two related texts
Figure 15: Comparison of the compact top-down CHESA representations of the words *smile* and *chord*

... dominant than the geographical context which results from the commonness of each of the contexts in language.

In figures 7 and 7 we show that when the word *bank* is put in context, CHESA resolves the ambiguity. Figure 7 shows the top-down representation of size 20 for the phrase "bank deposit". We see that all the concepts in this representation are related to the financial meaning of the word. Figure 7 shows the top-down representation of size 20 for the phrase "river bank". We can see that the concepts in the representation are related to the geographical context of the word.

4.6 Semantic Relatedness vs. Semantic Similarity

In Sections 4.2 and 4.3 we evaluated CHESA on the semantic relatedness task. A related, but different task, is computing semantic similarity between words and texts. Semantic relatedness is a more general notion than semantic similarity (Budanitsky et al., 2006) as words can be related in many different ways. While similar words are also related, the opposite is not always true. For example, while *(midday, noon)*, *(cock, rooster)* or...
Figure 16: Comparison of the compact top-down CHESA representations of the words *death* and *row*

*gem, jewel* are pairs of words which are both *similar* and *related*, *(headache, neurology)* or *(Obama, Washington)* are considered *related* by not *similar*.

CHESA algorithms assess semantic relatedness of words and texts. For example, Figure 4.6 shows the CHESA representations of the words *headache* and *neurology* that share several concepts. Therefore, we hypothesize that CHESA performance will degrade on semantic similarity tasks, as it will disagree with human judgements in case of related, but not similar text pairs.

To test our hypothesis we evaluate CHESA on two additional benchmarks for assessment of semantic similarity rather than semantic relatedness. The first benchmark, introduced by Li et al. (Li et al., 2006) consists of a collection of 65 sentence pairs that correspond to the Rubenstein and Goodenough (R&G) (Rubenstein & Goodenough, 1965) list of 65 word pairs. Each word from the R&G data set was replaced with its definition from the Collins Cobuild dictionary (Sinclair, 2001). The sentence pairs were ranked by 32 volunteers, all
native English speakers and graduates, and the average scores of their judgements was assigned to the pairs. As the distribution of the similarity scores assessed by the judges was heavily skewed towards low similarity scores, a subset of 30 pairs was selected to preserve more uniform distribution of scores (Li et al., 2006).

The second benchmark is the Microsoft paraphrase corpus (Dolan, Quirk, & Brockett, 2004). This benchmark consists of pairs of texts, collected for a period of 18 months, from thousands of news resources on the Web. The pairs were assigned binary labels by two human judgements, indicating whether the sentences are semantically equivalent paraphrases or not. The collection consists of a training set of 4,076 texts pairs, and a test set of 1,725 text pairs.

On the Li benchmark, CHESA obtains a Pearson’s correlation coefficient of 0.87 with human judgements\(^{19}\). These results slightly outperform previous work (Islam & Inkpen, 2008; Li et al., 2006) although the difference is not statistically significant.

\(^{19}\) Using the union bottom-up algorithm for \(k = 200\). We conducted experiments with \(k = 100, 300, 400\) and achieved similar results.
Figure 18: Compact top-down CHESA representation of the word bank

Figure 19: Compact top-down CHESA representation of the text "bank deposit"
On the Microsoft paraphrase benchmark the top-down CHESA algorithm obtains an F-Measure of 79.6 and the bottom-up CHESA algorithm obtains an F-measure of 79.7. The similarity threshold (0.3) was selected using the training set. Previous methods (Islam & Inkpen, 2008; Corley & Mihalcea, 2005; Mihalcea, Corley, & Strapparava, 2006) achieved slightly higher F-measures of 81.2-81.3.

According to the results on the Li benchmark, our hypothesis ought to be rejected. However, a further examination of the Li collection revealed that it mostly contains either similar text fragments or, neither similar nor related text fragments. This serves an explanation as to why CHESA performs well on this dataset. The Microsoft paraphrase corpus, on the other hand, contains many pairs of texts which are highly related but are not paraphrases. For example, the text “Ballmer has been vocal in the past warning that Linux is a threat to Microsoft” is very related to the text “In the memo, Ballmer reiterated the open-source threat to Microsoft”, although they are not considered as semantically equivalent paraphrases. Indeed, CHESA performance on this dataset is inferior comparing to methods that assess semantic similarity.

5. Related Work

Many statistical approaches have been introduced in the past for representing word and text semantics and assessing semantic relatedness.

The most common statistical approach is representing a text as a bag-of-words (BOW), namely as the collection of its words. Texts are compared in the vector space model (Salton & McGill, 1986; Baeza-Yates & Ribeiro-Neto, 1999), where they are represented as vec-

---

20. For $k = 200$. Similar results were obtained with $k = 100, 300, 400$
Figure 21: Comparison of the compact top-down CHESA representations of the words *neurology* and *headache*

tors in the hyperspace defined by single words. Semantic relatedness is then assessed by measuring the distance between these vectors. Although widely applied, this technique is inherently unable to compare single words, and performs sub-optimally on texts using different vocabularies to describe related topics.

Due to the limitations of the BOW approach, many methods began exploiting external knowledge to represent semantics and to assess semantic relatedness of words and texts.
Lexical resources such as WorldNet (Fellbaum, 1998) and Roget’s Thesaurus (Roget, 1852) have been utilized by many previous methods for semantic representation and semantic relatedness assessment. Most of these methods represent an input word with a node in the semantic graph defined by these lexical resources. Semantic relatedness is then computed using various metrics. Some methods use path length between the corresponding nodes (Lee et al., 1993; Rada, Mili, Bicknell, & Blettner, 1989; Leacock & Chodorow, 1998), and the information content (Resnik, 1995, 1999; Lin, 1998; Jiang & Conrath, 1997; Seco, Veale, & Hayes, 2004; Budanitsky et al., 2006) or the depth of the lowest common ancestor (Wu & Palmer, 1994). Others (Hughes & Ramage, 2007; Wan & Peng, 2005), perform random walks on the graph or use the Earth Movers distance (Rubner et al., 2000) to compute semantic distance. Finally, some techniques use a combination of different lexical resources (Li, Bandar, & McLean, 2003; Li et al., 2006; Mihalcea & Corley, 2006; Rodriguez & Egenhofer, 2003).

The methods that use these lexical resources have several limitations. First, the resources encode lexical information but lack comprehensive world knowledge. The semantic representation of an input word is essentially its mapping to a specific node, which does not allow deep interpretation of its semantics. Moreover, these resources are limited in terms of vocabulary as they only contain nouns, verbs, adjectives and adverbs. In the absence of information regarding proper names, neologisms, domain-specific terms etc., the vocabulary of words that can be semantically represented is very restricted. Finally, the aforementioned methods are limited to representing single words, while representing and comparing texts requires an extra level of sophistication (Mihalcea et al., 2006; Corley & Mihalcea, 2005; Islam & Inkpen, 2008).

Word-text co-occurrence data gathered from an unlabeled textual corpora is another resource of external knowledge for some semantic representations methods (Lund & Burgess, 1996; Landauer & Dumais, 1997; Foltz, 1996). One approach that uses such data is Latent Semantic Analysis (LSA) (Deerwester et al., 1990; Landauer & Dumais, 1997; Rehder et al., 1998; Wolfe et al., 1998; Landauer et al., 1998) which is based on singular value decomposition (SVD). LSA represents meaning of texts as a vector in a hyperspace of "latent concepts" retrieved by performing SVD on the co-occurrence matrix. Semantic relatedness is assessed by comparing the representations in the latent space.

Retrieving latent concepts or topics to describe text semantics was also addressed by probabilistic approaches. The Probabilistic Latent Semantic Analysis (PLSA) method (Hofmann, 1999) proposes a generative probabilistic model. According to this model, each document postulates a probability distribution on a latent multinomial random variable that represent a set of latent topics. The words of the document are a sample from a mixture model defined by this multinomial random variable. The Expectation Maximization algorithm is used to compute the conditional posterior probability distribution of the latent variable given a document. Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) is another probabilistic approach. The difference between this approach and PLSA is that the former assigns Dirichlet priors to the topics. Hierarchical Latent Dirichlet Allocation (hLDA) method (Blei, Griffiths, Jordan, & Tenenbaum, 2004) proposes an extension of LDA in which the latent topics lie in a hierarchy.

Both LSA and the aforementioned probabilistic methods represent words and text using "latent concepts". These concepts are extracted using SVD or Bayesian inference from
a corpus of training documents and are hard to interpret, as they cannot be mapped to knowledge concepts that are clear to humans. CHESA on the other hand represents meaning explicitly through natural concepts that are directly defined by humans and represent world knowledge. In addition, CHESA leverages the explicit hierarchical inner structure of those concepts encoded in Wikipedia categories. This is as opposed to the hierarchical LDA that learns the hierarchy as part of the inference process.

Web search engines have been exploited by several methods to compute semantic similarity. Some methods leverage search results returned for words and phrase queries to assess their semantic relationships (Danushka et al., 2007; Matsuo et al., 2006; Chen et al., 2006; Sahami & Heilman, 2006), while others (Maguitman et al., 2006), compare web pages using their organization under topical ontologies such as the The Open Directory Project and using graph-based metrics. CHESA differs from these techniques as it is able to process arbitrary texts and is not limited to existing Web pages or short texts that can constitute search queries.

Some methods (Lakkaraju et al., 2008; Gabrilovich & Markovitch, 2007a, 2005) map documents into hierarchical taxonomies, such as the Open Directory Project. These techniques are similar to CHESA in the sense that they use a pre-defined hierarchical ontology, and represent documents as hierarchical structures within that ontology. Nevertheless there are several substantial differences between CHESA and these methods.

CHESA generates balanced representations of words and texts, capturing their semantic gists on a high level of abstraction when the representations is compact, and generating more specific concepts as size limitations loosen. The aforementioned methods, on the other hand, primarily detect the concepts which are highly associated with the input word or text (e.g., measure the TFIDF of the input word or text in the text of the concept). This generation process triggers often overly-specific concepts usually located deep within the hierarchy. The algorithms then proceed by adding all the ancestors of the generated concepts to the representation, which results in representations having a small number of long paths from the root. Figure ?? shows an illustration of the typical structure of compact representations generated by these methods vs. representations generated by CHESA. Moreover, as opposed to CHESA, these methods are unable to control neither the final size of the representations nor their level of abstraction, as they select concepts regardless of their location in the hierarchy. Finally, these algorithms cannot manage redundancies. For instance if a parent and a child categories are very similar, and are highly associated with the input word or text, they both will be part of the semantic representation. CHESA refrains from such redundancies by using the conditional over-representation criterion.

CHESA is not the first methodology that uses Wikipedia to represent semantics and compare words and texts. WikiRelate! (Strube & Ponzetto, 2006) represents a word as a short list of Wikipedia articles containing the word in their title, and computes semantic relatedness by either comparing the texts of the corresponding articles or computing the path lengths or using information content for these articles in the Wikipedia category hierarchy. As opposed to CHESA, this method is limited to representing words only, and its vocabulary is bounded to terms appearing in titles of Wikipedia articles. Note that the method’s usage of Wikipedia category graph is different than ours. While CHESA uses Wikipedia conceptual hierarchy for generating the semantics of text (regardless of the spe-
specific application), WikiRelate! uses Wikipedia categories only for computing the semantic distance between words and texts.

Wikipedia categories have also been used for document topic identification (Schonhofen, 2006), where documents are classified to Wikipedia categories according to the titles of the articles in those categories. Other methods (Milne, 2007; Kaiser et al., 2009) utilize the hyperlink structure of Wikipedia to compute semantic relatedness of words and texts.

Explicit Semantic Analysis (ESA) (Gabrilovich & Markovitch, 2006, 2007b, 2009) represents semantics of words and texts as a flat weighted vector of Wikipedia articles, where the weights are assigned using a TFIDF metric between the text represented and the textual content of the Wikipedia article. We have extensively described and illustrated the differences between CHESA and ESA throughout this paper.

There is a certain connection between hierarchical semantic representation of texts based on a global conceptual hierarchy and hierarchical classification of texts. These methods (e.g. (Koller & Sahami, 1997; Dumais & Chen, 2000; Ruiz & Srinivasan, 2002)), classify documents to one node in a topic hierarchy by a multiple-phase process where a classifier is built for every set of siblings in the hierarchy. Namely, a chain of classifiers is applied to the input text, classifying it to a more specific topic each iteration. The local classification can be performed with a variety of classifiers such as Bayesian classifiers and Support Vector Machines. Although making local decisions within a topical hierarchy and able to classify a document to topics with alternating abstraction level, hierarchical classification is essentially a mapping of a document to a single node within the topical hierarchy rather than an hierarchical representation which constitutes a sub-hierarchy of the global hierarchical ontology.

6. Discussion

Researchers has recognized, long time ago, that many natural language processing tasks require comprehensive semantic interpretation of text. The efforts to find an appropriate semantic representation with associated interpretation method, have followed two main directions. One direction, best exemplified by the CYC project (Lenat & Guha, 1989), use extensive manual efforts to construct rich and deep semantics. The other direction, exemplified by Latent Semantic Analysis (Landauer & Dumais, 1997), automatically generate semantics based on statistical analysis of large text collection. The generated semantics, however, are usually not appropriate for deep natural language understanding.

The CHESA framework presented in this paper, tries to bridge between these two approaches. On one hand the semantics is generated automatically. On the other hand, the generated representation is structured, and coincides with human conceptual hierarchy, thus having the potential of being used for deep NLP.

CHESA representation is flexible with respect to size and level of abstraction. Even when compact, the representation is powerful and expressive. Since CHESA uses Wikipedia as a resource of world knowledge, the rapid growth of the online encyclopedia directly improves the richness of the semantic representations.

We developed a methodology to compute semantic relatedness based on our representation and showed that CHESA is superior to previous methods when representations are
compact. For unlimited representation size, empirical analysis showed that CHESA is comparable to ESA and is superior to other methods.

We intend to expand our study of CHESA and design methods for using CHESA for additional text processing tasks, such as word sense disambiguation, text categorization and information retrieval. We believe that the novel method introduced in this paper is a step forward in making statistical semantic representation methods useful for deep natural language understanding.

References


Appendix A. Building a Conceptual Hierarchy from Wikipedia Concept Graph

Algorithm 3 describes the procedure of constructing a rooted directed acyclic graph of concepts from the Wikipedia graph of categories and articles. The graph is processed in a bread-first-search order and paths leading from the root to concepts that can be reached by shorter paths are removed. Moreover, concepts that cannot be reached from the root are not part of the final hierarchy.

\textbf{Algorithm 3 Build Conceptual Hierarchy}

\begin{itemize}
  \item \textbf{Input:} A directed graph $G(V, E)$, a root node $v_r \in V$
  \item \textbf{Output:} A multiple-inheritance hierarchy $G'(V', E')$ (a rooted directed acyclic graph)
  \item 1: $V' \leftarrow \emptyset$
  \item 2: $E' \leftarrow \emptyset$
  \item 3: $Q \leftarrow$ an empty First In First Out queue
  \item 4: \textbf{for all} $v \in V$ \textbf{do}
  \item 5: $v\.depth = 0$
  \item 6: \textbf{end for}
  \item 7: $Q\.add(v_r)$
  \item 8: \textbf{while} $Q$ is not empty \textbf{do}
  \item 9: $v \leftarrow Q\.getFirst()$
  \item 10: $Q\.removeFirst()$
  \item 11: $V' \leftarrow V' \bigcup \{v\}$
  \item 12: \textbf{for all} $v'$ such that $(v, v') \in E$ \textbf{do}
  \item 13: \textbf{if} $((v'.depth = 0) \text{ or } (v'.depth = v\.depth + 1) \text{ or } (v' \text{ is a leaf}))$ \textbf{then}
  \item 14: $E' \leftarrow E' \bigcup \{(v, v')\}$
  \item 15: $v'.depth \leftarrow v\.depth + 1$
  \item 16: $Q\.add(v')$
  \item 17: \textbf{end if}
  \item 18: \textbf{end for}
  \item 19: \textbf{end while}
  \item 20: \textbf{return} $V', E'$
\end{itemize}

Appendix B. Top-Down vs. Bottom-Up Representation

Figures 22 and 23 show the top-down and bottom-up CHESA representations of 50 concepts for the word \textit{tiger}. 

46
Figure 22: The top-down representation if size 50 generated by CHESA for the word *tiger*
Figure 23: The bottom-up representation if size 50 generated by CHESA for the word tiger