Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Semantic Representation Using Explicit Concept Space Models

Walid Shalaby, Wlodek Zadrozny
Department of Computer Science
University of North Carolina at Charlotte
{wshalaby, wzadrozn}@uncc.edu

Abstract

Explicit concept space models have proven efficacy for text representation in many natural language and text mining applications. The idea is to embed textual structures into a semantic space of concepts which captures the main topics of these structures. Despite their wide applicability, existing models have many shortcomings such as sparsity and being restricted to Wikipedia as the main knowledge source from which concepts are extracted. In this paper we highlight some of these limitations. We also describe Mined Semantic Analysis (*MSA*); a novel concept space model which employs unsupervised learning in order to uncover implicit relations between concepts. *MSA* leverages the discovered concept-concept associations to enrich the semantic representations. We evaluate *MSA*'s performance on benchmark data sets for measuring lexical semantic relatedness. Empirical results show superior performance of *MSA* compared to prior state-of-the-art methods.

Several semantic representation models have been proposed in the literature. Most of them utilize textual corpora from which world knowledge is acquired and used to represent textual structures as high-dimensional "meaning" vectors.

Explicit concept space models such as *ESA* (Gabrilovich and Markovitch 2007) and *SSA* (Hassan and Mihalcea 2011) construct bag-of-concepts (*BOC*) vectors to represent textual structures (terms, phrases, documents) using concepts in encyclopedic knowledge source such as Wikipedia. Those *BOC* embeddings capture the main topics of the given text and therefore are useful for understanding its semantics.

The *BOC* representations have proven efficacy for semantic analysis of textual data especially short texts where contextual information is missing or insufficient. For example, measuring lexical semantic similarity/relatedness (Gabrilovich and Markovitch 2007), text categorization (Song and Roth 2014), search and relevancy ranking (Egozi, Markovitch, and Gabrilovich 2011), and others.

Existing concept space models such as *ESA* suffer many shortcomings. For example, the generated *BOC* vectors are sparse causing low similarity between texts in the concept space even when they are semantically similar. (Song and Roth 2015) proposed a densification mechanism by combining *ESA* and concept-concept similarity based on the neural embeddings of concept words. However, this mechanism disregards the fact that the concepts already have implicit associations/similarities which could be generated from their corresponding articles or from the structure of the employed knowledge base (e.g., its link graph).

Additionally, explicit concept models are restricted to Wikipedia as the main source of concepts. And it is unclear how domain specific knowledge could be utilized to generate domain specific conceptual representations.

Mined Semantic Analysis

We call our approach Mined Semantic Analysis (*MSA*) as it utilizes unsupervised data mining techniques in order to discover the concept space of a given textual structure. The motivation behind our approach is to mitigate a notable gap in prior concept space models which are limited to direct associations between texts and concepts. Therefore those models lack the ability to transfer the association relation to other latent concepts which contribute to the meaning of these texts.

In a nutshell, *MSA* generates the concept space of a given text by utilizing two repositories created offline: 1) the Wikipedia search index, and 2) the concept-concept associations repository. First, the explicit concept space is constructed by retrieving concepts (titles of articles) explicitly mentioning the given text. Second, latent concepts associated with each of the explicit concepts are retrieved from the associations repository.

Concept-Concept Association Mining

We employ rule mining (Agrawal, Imieliński, and Swami 1993) in order to learn implicit concept-concept associations using the *"See also"* link graph of Wikipedia.

Formally, given a set of concepts $C = \{c_1, c_2, ..., c_N\}$ of size *N* (i.e., all Wikipedia article titles). We build a dictionary of transactions $T = \{t_1, t_2, t_3, ..., t_M\}$ of size *M* such that $M \leq N$. Each transaction *t* in *T* contains a subset of concepts in *C*. *t* is constructed from each article in Wikipedia that contains at least one entry in its *"See also"* section. For example, if an article representing concept c_1 with entries in its *"See also"* section referring to concepts $\{c_2, c_3, ..., c_n\}$, a transaction $t = \{c_1, c_2, c_3, ..., c_n\}$ will be created and added to *T*. A set of rules *R* are then learned by mining *T*. Each rule *r* in *R* is defined as

$$r(s, f) = \{(X \Rightarrow Y) : X, Y \subseteq C \; and \; X \cap Y = \emptyset\} \quad (1)$$

Both *X* and *Y* are subsets of concepts in *C*. Rule *r* is parameterized by two parameters: 1) Support (*s*) indicating how many times both *X* and *Y* appeared together in *T*, and 2) Confidence (*f*) denoting *s* divided by number of times *X* appeared in *T*. After learning *R*, we end up having concept(s)-concept(s) associations. Using such rules, we can experiment with the strength of those associations as a function of *s* and *f*.

Constructing the Concept Space

Given a set of concepts *C* of size *N*, *MSA* constructs the bag-of-concepts vector C_t of term(s) *t* through two phases: search and expansion. In the search phase, *t* is represented as a search query and is searched for in the Wikipedia search index. This returns a weighted set of concepts C_s that best match *t* based on the vector space model. In the expansion phase, we use inferred association rules to expand each concept *c* in C_s by looking for its associated set of concepts in *R*. Formally, the expansion set of concepts C_p is obtained by

$$C_p = \bigcup_{c \in C_s, c' \in C} \{(c', w) : \exists r(s, f) = c \Rightarrow c'\} \quad subject\ to: \ s >= \epsilon, f >= \upsilon \tag{2}$$

Note that we add all the concepts that are implied by *c* where this implication meets the support and confidence thresholds (ϵ, υ). The weight of c' is denoted by *w*; currently we use simple weight propagation mechanism where all concepts implied by *c* inherit the same weight as *c*. Finally, all obtained concepts from search and expansion phases are merged to construct the concept vector C_t of term(s) *t*.

Experiments on Semantic Relatedness

We evaluate *MSA*'s performance on benchmark data sets for measuring lexical semantic relatedness[1]. Each data set is a collection of word pairs along with human judged similarity/relatedness score for each pair. We use the *MC* data set containing 30 pairs, the *RG* data set containing 65 pairs, the *WS* data set containing 353 similar/related pairs, and the *WSS* & *WSR* data sets containing similar & related pairs from the *WS* data set respectively.

We report the results by measuring Spearman rank-order correlation (ρ) between *MSA*'s computed relatedness scores and the gold standard provided by human judgments. We compare our results with those obtained from two types of semantic representation models. First, explicit semantics models such as *ESA* (Gabrilovich and Markovitch 2007) and *SSA* (Hassan and Mihalcea 2011). Second, the modern neural language models such as *Word2Vec* (Baroni, Dinu, and Kruszewski 2014).

Table 1 shows *MSA*'s Spearman correlation scores compared to other models on the five data sets. As we can see, *MSA* gives the highest correlation scores on all data sets. These results demonstrate *MSA*'s potential for augmenting the explicit concept space by other semantically related concepts which contribute to understanding the given text.

[1]https://github.com/mfaruqui/eval-word-vectors/tree/master/data/word-sim

	MC	*RG*	*WSS*	*WSR*	*WS*
ESA$^{\diamond}$	0.83	0.80	0.74	0.68	0.72
SSA$_s$*	0.81	0.83	–	–	0.63
SSA$_c$*	0.84	0.83	–	–	0.60
Word2Vec$^{\triangleright}$	0.82[2]	0.84	0.76	0.64	0.71
MSA	**0.87**	**0.86**	**0.77**	**0.71**	**0.73**

Table 1: *MSA*'s Spearman (ρ) scores on benchmark data sets vs. other techniques. ($\diamond$) our implementation, ($\star$) from (Hassan and Mihalcea 2011), and ($\triangleright$) from (Baroni, Dinu, and Kruszewski 2014).

Moreover, it is clear that augmenting the *BOC* vectors with implicitly related concepts achieves performance gains over other explicit concept representations such as *ESA* and *SSA*. (Shalaby and Zadrozny 2015) provide more details about these results and an analysis of their statistical significance.

Conclusion

In this paper, we presented *MSA*, a novel explicit concept space model which employs unsupervised data mining techniques to create conceptual semantic representations of text.

MSA is motivated by inability of prior concept space models to capture implicit relations between concepts. To this end, *MSA* mines for implicit concept-concept associations through Wikipedia's *"See also"* link graph. We demonstrated through empirical results the effectiveness of *MSA*'s representation for measuring lexical semantic similarity.

References

Agrawal, R.; Imieliński, T.; and Swami, A. 1993. Mining association rules between sets of items in large databases. In *ACM SIGMOD Record*, volume 22, 207–216. ACM.

Baroni, M.; Dinu, G.; and Kruszewski, G. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, 238–247.

Egozi, O.; Markovitch, S.; and Gabrilovich, E. 2011. Concept-based information retrieval using explicit semantic analysis. *ACM Transactions on Information Systems (TOIS)* 29(2):8.

Gabrilovich, E., and Markovitch, S. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, volume 7, 1606–1611.

Hassan, S., and Mihalcea, R. 2011. Semantic relatedness using salient semantic analysis. In *AAAI*.

Shalaby, W., and Zadrozny, W. 2015. Measuring semantic relatedness using mined semantic analysis. *arXiv preprint arXiv:1512.03465*.

Song, Y., and Roth, D. 2014. On dataless hierarchical text classification. In *AAAI*, 1579–1585.

Song, Y., and Roth, D. 2015. Unsupervised sparse vector densification for short text similarity. *Proc. North Am. Chapter Assoc. Computat. Linguistics* 1275–1280.