Ensemble similarity measures for clustering terms

Ittoo, Ashwin; Maruster, Laura

Published in:
2009 WRI World Congress on Computer Science and Information Engineering, CSIE 2009

DOI:
10.1109/CSIE.2009.764

Ensemble Similarity Measures for Clustering Terms

Ashwin Ittoo, Laura Maruster
Faculty of Economics and Business
University of Groningen
9700 AV Groningen, The Netherlands
r.a.ittoo@rug.nl

Abstract

Clustering semantically related terms is crucial for many applications such as document categorization, and word sense disambiguation. However, automatically identifying semantically similar terms is challenging. We present a novel approach for automatically determining the degree of relatedness between terms to facilitate their subsequent clustering. Using the analogy of ensemble classifiers in Machine Learning, we combine multiple techniques like contextual similarity and semantic relatedness to boost the accuracy of our computations. A new method, based on Yarowsky's [9] word sense disambiguation approach, to generate high-quality topic signatures for contextual similarity computations, is presented. A technique to measure semantic relatedness between multi-word terms, based on the work of Hirst and St. Onge [2] is also proposed. Experimental evaluation reveals that our method outperforms similar related works. We also investigate the effects of assigning different importance levels to the different similarity measures based on the corpus characteristics.

1. Introduction

It is not uncommon for different terms or jargons to be used as the linguistic realization of a single concept. For example, "surface" and "glass" could represent the concept "screen". In domains like bio-technology, locating newly coined terms, not yet formally encoded, may be challenging. There is a pressing need in knowledge-intensive domains for innovative techniques to systematically structure information; typically available as textual documents, and represented by corresponding terms. One way to structure knowledge is by clustering similar terms based on the degree to which they are semantically related.

Identifying semantically related terms is an indispensable step in numerous applications like document categorization, word sense disambiguation, and ontology learning [6], among others.

In this research, we present an ensemble term similarity computation method incorporating the contextual similarity and semantic relatedness measures. Most semantic relatedness measures rely on the relative positions of terms within the WordNet's [3] concept hierarchy [7]. They are unable to deal with terms that are not defined within WordNet, and with multi-word terms. Our approach supplements the semantic relatedness between two terms with their corresponding contextual similarity. These two measures, applied in an ensemble, complement each other's limitations and boost the accuracy of similarity computations. Contextual similarity reduces the reliance on WordNet, and is computed from topic signatures by adapting Yarowsky's word sense disambiguation approach [9]. We modified the semantic relatedness measure of Hirst and St. Onge [2] to deal with multi-word terms efficiently.

2. Related work

2.1. Contextual similarity

The contextual similarity between two terms is based on the number of surrounding context words they have in common. It has been employed in word sense disambiguation [9], enriching ontologies [1], and term similarity computations [5].

2.2. Lexical similarity

Multi-word terms are lexically related if they share a common head [4, 5]. In [4], taxonomic relations between terms were identified based on shared nouns of lexically similar terms. Nenadić et al. [5] measured

978-0-7695-3507-4/08 $25.00 © 2008 IEEE
DOI 10.1109/CSIE.2009.764

the lexical similarity between terms from their shared head noun phrases and additional modifiers.

2.3. Semantic relatedness

Semantic relatedness defines the degree of relatedness between two lexically expressed concepts [7]. It is more general than semantic similarity, and takes into account different relations (hyponymy, meronymy and others) between seemingly dissimilar terms. Most semantic relatedness measures compare terms [7] based on their relative positions and on the path lengths separating them in WordNet. Hence, they are restricted to terms within WordNet, and may not perform well on multi-word terms.

2.4. Combining similarity measures

Nenadić et al. [5] presented a methodology incorporating contextual, lexical and syntactic similarity measures. Contextual similarity was defined as the ratio of common to distinct context patterns. Lexically similar terms were identified based on their common head nouns. Syntactically similar terms were those co-occurring in certain lexico-syntactic patterns. However, these patterns are heavily corpus-dependent, and not reliable for measuring similarity. It was also found that none of the similarity measures were very reliable on their own, and had to be combined for improved performance.

3. Proposed methodology

Our proposed methodology for term similarity computation incorporates the contextual similarity and semantic relatedness measures between terms. We excluded syntactic similarity due to its well-reported limitations. Semantic relatedness, which considers different relations between terms, implicitly encompasses semantic and lexical similarities, and alleviates the need for their separate evaluation.

When applied together, contextual similarity and semantic relatedness measures complement each other's shortcomings. Contextual similarity reduces the reliance on WordNet that affects semantic relatedness, while semantic relatedness compensates for low-quality contexts that may affect contextual similarity. This leads to the accurate identification of related terms, and to the subsequent generation of high-quality clusters.

Our contextual similarity measure adapts Yarowsky's word sense disambiguation procedure [9]. We generate high-quality topic signatures from terms' contexts in order to compute the contextual similarity. We modified the semantic relatedness measure of Hirst and St. Onge [2] to handle multi-word terms efficiently.

3.1. Proposed contextual similarity measure

Each term *t* is represented by its feature vector. In our case, this corresponds to the topic signature, *TS(t),* of term *t*.

$$TS(t)=\{\langle w_1, s_1\rangle,\ldots\langle w_i, s_i\rangle,\ldots\langle w_n, s_n\rangle\}$$

Each tuple of *TS(t)* consists of a word w_i that co-occurs with *t*, and of s_i, which is the mutual information (MI) between w_i and *t*. The value of s_i is given in [9] as $\log\frac{P(w_i\,|\,t)P(t)}{P(w_i)}$; where $P(w_i\,|\,t)$ is the probability of w_i appearing in *t*'s context, $P(t)$ is the probability of *t*'s occurrence, $P(w_i)$ is the probability of w_i's occurrence. Computing the MI ensures that only the most relevant context words are selected. The contextual similarity between terms *t1* and *t2* is the cosine similarity between their topic signatures *TS(t1)* and *TS(t2)* respectively, as shown in equation (1)

$$ConSim(t1,t2)=\frac{TS(t1)\bullet TS(t2)}{|TS(t1)||TS(t2)|} \qquad (1)$$

3.2. Proposed semantic relatedness measure

Unlike other semantic relatedness measures, the approach of Hirst and St. Onge [2] is not restricted to nouns and hyponymy relations. Furthermore, it implicitly encompasses semantic similarity and lexical similarity [7]. It is calculated as in equation (2)

$$SemR(t1,t2)=C-path(t1,t2)-(k\times d) \qquad (2),$$

where *path* is the path length between terms *t1* and *t2*, *d* is the number of direction changes, *C* and *k* are constants.

Due to its reliance on WordNet, this method may perform poorly when measuring the relatedness between multi-word terms. We now describe a novel approach to alleviate this problem.

We treat each term as a "bag of words". Given $t1=\langle w_{1,1}, w_{1,2}, w_{1,3}\rangle$ and $t2=\langle w_{2,1}, w_{2,2}\rangle$, where $w_{i,j}$ is the ith word of term *tj*. A recursive algorithm, depicted in Figure 1, is then followed.

Step1: Find the cross product between the words of *t1* and *t2*. This yields a set of tuples.

$$(t1 \otimes t2) = \begin{Bmatrix} \langle w_{1,1}, w_{2,1} \rangle, \langle w_{1,2}, w_{2,1} \rangle, \langle w_{1,3}, w_{2,1} \rangle, \\ \langle w_{1,1}, w_{2,2} \rangle, \langle w_{1,2}, w_{2,2} \rangle, \langle w_{1,3}, w_{2,2} \rangle \end{Bmatrix}$$

Step2: Treat each word in the tuples as a term. Compute the semantic relatedness between pairs of words in the tuples using equation (2). That is,

$\forall \langle w_x, w_y \rangle \in (t1 \otimes t2) : SemR(w_x, w_y)$.

For simplicity, a tuple is shown as $\langle w_x, w_y \rangle$.

Step3: Let the tuple with the most semantically related pair of words be $\langle w_x, w_y \rangle$.

Accumulate the relatedness measure between the words of $\langle w_x, w_y \rangle$:

$$tempSem += \arg\max_{\langle w_x, w_y \rangle} SemR(w_x, w_y)$$

Remove tuple $\langle w_x, w_y \rangle$ from $(t1 \otimes t2)$.

Step4: Repeat Steps 2 and 3 until $(t1 \otimes t2)$ is empty.

Figure 1. Semantic relatedness for multi-word terms

Taking the cross-product of the terms' constituent words in Step1 ensures the identification of the most semantically related pair of words even if they appear in different order in the terms. After the procedure of Figure 1 is run, *tempSem* indicates the semantic relatedness between *t1* and *t2*. Since terms of the same length intuitively exhibit greater relatedness than those of different lengths, terms with different lengths are penalized by normalizing their relatedness score with the ratio of their lengths

$$NormSemR(t1, t2) = tempSem \times \frac{\min(|t1|, |t2|)}{\max(|t1|, |t2|)}$$

, where $|tj|$ is the length of term *tj*.

3.3. Ensemble similarity measure

We define our ensemble similarity measure between terms *t1* and *t2*, *ES(t1,t2)*, as a combination of the contextual similarity and of the normalized semantic relatedness measures between *t1* and *t2*.

$$ES(t1, t2) = ConSim(t1, t2) + NormSemR(t1, t2)$$

4. Experiments

To demonstrate the accuracy and robustness of our method, evaluation was performed on completely unstructured, noisy, free-text downloaded from the Internet, as opposed to most previous works where evaluation involved highly-domain specific and (semi-) structured corpora. Our evaluation corpus consisted of mobile phone descriptions from vendors' sites and customers' opinions from online forums. The corpus' free-text nature presented additional challenges; such as identifying syntactic dependencies between words in a term's context to form the topic signatures, and dealing with significant noise level.

4.1. Generating target terms

The downloaded documents are pre-processed (parsed, cleansed from stop-words and noise), to result in a corpus of 500 documents, each with an average of 15 sentences. Relevant terms from the corpus, whose TF-IDF [8] weights exceeded an experimental threshold, were selected as target terms to be subsequently clustered. We considered only terms with at most 3 constituent words.

4.2. Generating topic signatures

The topic signature of a term *t* is the set of contextual words that co-occur in its surrounding, together with a measure indicative of their frequencies or saliency. For the generation of high-quality topic signatures, all contextual words are represented by their Part-of-Speech (POS) tags. For example, the contextual word (adjective) *large* co-occurring with the target term *display* (…*large display*…) is replaced by *JJ.* Furthermore, with syntactic parsing, only those contextual words that exhibit a dependency relation with target terms are considered for further analyses. Then, concordances of the 5 most salient ones are extracted from either side of the target term (context width of 10). The saliency of a word (or its POS tag) with respect to a target term is its mutual information (MI) [9] with the target term. Table 1 lists sample topic signatures for the target term *display.*

Table 1. Topic signatures

Context word/POS Tag	Mutual Information
Large /JJ	*0.53*
Pixels/NNS	*0.76*
Color/NNS	*0.62*

4.3. Measuring term similarity

Following our approach for ensemble term similarity computation, described in Section 3, a similarity matrix is created. It lists the degree to which the identified target terms are related to one another based on their contextual similarity and on their semantic relatedness. Each matrix row represents a similarity vector corresponding to a specific target term. A sample similarity vector, specifying the degree to which the multi-word term *LCD Display* is related to other terms as measured by our proposed approach is illustrated in Table 2. Calculated similarity values are shown in brackets.

Table 2. Similarity vector

LCD Screen (0.83)	Color Screen (0.75)	TFT Color Screen (0.73)	…	Camera (0.51)

To illustrate the strength of our approach, we note its correct identification of the strong relation between *TFT Color Screen* and *LCD Display*, although they do not exhibit any apparent resemblance, are multi-word terms not defined within WordNet, and are of different lengths. However, our method also suggests *Camera* as being closely related to *LCD Display*, which is not correct. This could be attributed to the large number of common contextual words (*high-resolution*, *pixels*,…) shared by both *Camera* and *LCD Display*.

4.4. Clustering

The similarity matrix created in the previous stage is fed to clustering algorithms. We investigated both k-Means and hierarchical clustering approaches, but report only the latter's results due to higher observed performance during the experiments. We adopted an agglomerative hierarchical clustering approach that successively merges pairs of clusters, until a single cluster encompassing all previously generated clusters is created. Clusters are fused following the Ward procedure, which aims to minimize intra-cluster variance.

By cutting off the resulting dendogram at a particular depth, a set of around 50 clusters were obtained, with sizes varying from 2 to 15 terms, singletons excluded.

Our method to determine the precision is briefly described next. We randomly selected 14 candidate clusters, each of different size (ranging from 2 to 15). The clustering is reformulated as a classification task, with each cluster corresponding to a category and terms within each cluster as instances classified under that class. For example, consider cluster C_3 (size 3) with terms *{LCD Display, TFT Color Screen, Camera}*. Since the Ward's procedure generates relatively small clusters (maximum size 15), their visual/manual inspection for determining the precision was possible. The majority of terms in C_3 deal with *Display/Screen*, and hence *Camera* can be considered as a “ false positive”. The precision P_3 of C_3 is then calculated as

$$\frac{true\_positive}{true\_positive + false\_positive} = \frac{2}{3} = 67\%$$

The overall precision, $P_{overall}$ is obtained by averaging over the 14 clusters $P_{overall} = \frac{\sum_{i=2}^{15} P_i}{14} = 73\%$, where P_i is the precision of the cluster C_i of size *i*.

Despite our highly unstructured and free-text corpus, our precision of 73% is higher than that reported by [5], which achieved around 70% over a corpus consisting of bio-medical abstracts.

4.5. Introducing weights

Our proposed ensemble similarity measure is adaptable to suit different corpus characteristics by assigning different weights to its contextual similarity and its semantic relatedness components. When the corpus contains sufficient context to support target terms, a higher weight can be allocated to the contextual similarity measure. With sparse context, semantic relatedness can be given higher weight. We call our weighted ensemble similarity measure *wES*, and define it as

$$\alpha ConSim(t1, t2) + \beta NormSemR(t1, t2),$$

where α and β represent weights.

We repeated our experiments with the current corpus (mobile phone descriptions) but varying β, and with another corpus with sparse context (containing only technical specifications) but varying α. Results, in Table 3, indicate that when sufficient context is available, semantic relatedness does not contribute significantly to the overall similarity since precision drops by around only 12% (73% to 61%). However, when context is sparse, semantic relatedness alone cannot supplement contextual similarity, and the precision drops by around 26% (73% to 47%). This could be due to the over-reliance of semantic relatedness on WordNet, which causes it to fare poorly when confronted with terms not defined in WordNet. (The optimal values for α and β were tuned manually)

Table 3. Adjusting weights

Corpus	Max. Precision
Original (sufficient context)	61% (β=0.4, α=1)
Sparse context	47% (α=0.3, β=1)

5. Conclusion and future work

We presented a novel approach for measuring term relatedness by applying contextual similarity and semantic relatedness in an ensemble so that they complement each other. We applied our method to create a similarity matrix of target terms that were subsequently clustered using an agglomerative hierarchical clustering approach. We reported a clustering precision of 73%. By assigning different weights to the contextual similarity and to the semantic relatedness measures, our approach can be easily adapted to situations where sufficient context is available in the corpus or when context is sparse. However, experiments revealed that in the absence of context, semantic relatedness alone does not provide a reliable estimate of terms' similarity due to its over-reliance on WordNet.

The work presented in this paper could serve as a basis for concept grouping in ontology learning. It can even be extended for learning non-taxonomic relations between terms, which is an overlooked area in ontology learning. Further research can also be directed towards automatically determining the weights to be assigned to contextual similarity and to semantic relatedness based on the corpus characteristics.

6. References

[1] E. Agirre, O. Ansa, E. Hovy, and D. Martinez, "Enriching very large ontologies using the www", in *Proceedings of the Ontology Learning Workshop*, ECAI, Berlin, Germany, 2000.

[2] G. Hirst and D. St. Onge, "Lexical chains as representations of context for the detection and correction of malapropisms", in *WordNet: An electronic lexical database*, MIT Press, Cambridge, Massachusetts, 1998.

[3] G.A. Miller, "WordNet: A Lexical Database for English", in *Communications of the ACM*, 1995.

[4] D. I. Moldovan and R. Girju, "Domain-Specific Knowledge Acquisition and Classification Using WordNet", in *Proceedings of the Thirteenth International Florida Artificial Intelligence Research Society*, Florida, USA, 2000.

[5] G. Nenadić, I. Spasić, and S. Ananiadou, "Automatic discovery of term similarities using pattern mining ", in *COLING-02 on COMPUTERM 2002: 2nd International Workshop on Computational Terminology*, Taiwain, 2002.

[6] P. Pantel and D. Lin, "Discovering word senses from text", in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining,* Alberta, Canada, 2002.

[7] S. Patwardhan, S. Banerjee, and T. Pedersen, "Using measures of semantic relatedness for word sense disambiguation", in *Proceedings of the 4th International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, Mexico, 2003.

[8] G. Salton, and C. Buckley, "Term-weighting approaches in automatic text retrieval", *Information Processing & Management*, 1998, pp. 513-523.

[9] D. Yarowsky, "Word-Sense disambiguation using statistical models of Roget's categories trained on large corpora", in *Proceedings of the 14th Conference on Computational Linguistics*, Nantes, France. 1992.