Semantic Relatedness Using Salient Semantic Analysis

Samer Hassan and Rada Mihalcea
University of North Texas
Denton, Texas 76203
samer@unt.edu,rada@cs.unt.edu

Abstract

This paper introduces a novel method for measuring semantic relatedness using semantic profiles constructed from salient encyclopedic features. The model is built on the notion that the meaning of a word can be characterized by the salient concepts found in its immediate context. In addition to being computationally efficient, the new model has superior performance and remarkable consistency when compared to both knowledge-based and corpus-based state-of-the-art semantic relatedness models.

Introduction

Semantic relatedness is the task of finding and quantifying the strength of the semantic connections that exist between textual units, be they word pairs, sentence pairs, or document pairs. For instance, one may want to determine how semantically related are car and $automobile$, or $noon$ and $string$. To make such a judgment, we rely on our accumulated knowledge and experiences, and utilize our ability of conceptual thinking, abstraction, and generalization. Accordingly, a good system should not only be able to acquire and use a large amount of background knowledge, but it should also be able to abstract it and generalize it. To address this aspect, many semantic models have been introduced that integrate concept abstraction either explicitly (e.g., Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch 2007)), or implicitly (e.g., Latent Semantic Analysis (LSA) (Landauer et al. 1997)).

In this paper, we introduce a new model called Salient Semantic Analysis (SSA), which incorporates a similar semantic abstraction and interpretation of words, by using salient concepts gathered from encyclopedic knowledge.[1] The main idea underlying our method is that we can determine the semantic relatedness of words by measuring the distance between their concept-based profiles, where a profile consists of salient concepts occurring within contexts across a very large corpus. Unlike previous corpus-based methods of relatedness, which utilize word-word associations to create contextualized profiles, our model utilizes concepts that frequently co-occur with a given word. Moreover, we only use those concepts that have high saliency in a document, meaning that they are highly relevant to the given text. Through evaluations on standard benchmarks, consisting of word-to-word and text-to-text relatedness datasets, we show that the new SSA method exceeds the accuracy of previously proposed knowledge-based and corpus-based measures of relatedness.

[1] By "concept" we mean an unambiguous word or phrase with a concrete meaning, which can afford an encyclopedic definition.

Related Work

There are many approaches to semantic relatedness that have been proposed to date, and they can be generally grouped into two main categories: knowledge-based and corpus-based. Knowledge-based measures such as L&C (Leacock and Chodorow 1998), Lesk (Lesk 1986), Wu&Palmer (Wu and Palmer 1994), Resnik (Resnik 1995), J&C (Jiang and Conrath 1997), H&S (Hirst and St Onge 1998), and many others, employ information extracted from manually constructed lexical taxonomies like Wordnet (Fellbaum 1998), Roget (Jarmasz 2003), and Wiktionary (Zesch, Muller, and Gurevych 2008). While these methods show potential in addressing the semantic relatedness task, they are burdened by their dependence on static, expensive, manually constructed resources. Moreover, these measures are not easily portable across languages, as their application to a new language requires the availability of the lexical resource in that language.

On the other side, corpus-based measures such as LSA (Landauer et al. 1997), ESA (Gabrilovich and Markovitch 2007), Pointwise Mutual Information (PMI) (Church and Hanks 1990), PMI-IR (Turney 2001), Second Order PMI (Islam and Inkpen 2006), and distributional similarity (Lin 1998) employ probabilistic approaches to decode the semantics of words. They consist of unsupervised methods that utilize the contextual information and patterns observed in raw text to build semantic profiles of words. While most of these corpus-based methods induce semantic profiles in a word-space, where the semantic profile of a word is expressed in terms of their co-occurrence with other words, ESA and LSA stand out as different, since they rely on a concept-space representation. In these two methods, the semantic profile of a word is expressed in terms of the explicit (ESA) or implicit (LSA) concepts. This departure from

the sparse word-space to a denser, richer, and unambiguous concept-space resolves one of the fundamental problems in semantic relatedness, namely the vocabulary mismatch. These concept-based approaches are powerful and competitive when compared to the knowledge-based measures. They are also scalable due to their unsupervised nature. One of the methods closely related to our work is ESA (Gabrilovich and Markovitch 2007), which uses encyclopedic knowledge in an information retrieval framework to generate a semantic interpretation of words. Since encyclopedic knowledge is typically organized into concepts (or topics), each concept is further described using definitions and examples. ESA relies on the distribution of words inside the encyclopedic descriptions. It builds semantic representations for a given word using a word-document association, where the document represents a Wikipedia article (concept). In this vector representation, the semantic interpretation of a text fragment can be modeled as an aggregation of the semantic vectors of its individual words. Also closely related is the Latent Semantic Analysis (Landauer et al. 1997) model (LSA). In LSA, term-context associations are captured by means of a dimensionality reduction operated by a singular value decomposition (SVD) on the term-by-context matrix $\mathbf{T}$, where the matrix is induced from a large corpus. This reduction entails the abstraction of meaning by collapsing similar contexts and discounting noisy and irrelevant ones, hence transforming the real world term-context space into a word-latent-concept space which achieves a much deeper and concrete semantic representation of words.

Salient Semantic Analysis

We derive semantic profiles based on the Wikipedia corpus by using one of its most important properties – the linking of concepts within articles. The links available between Wikipedia articles, obtained either through manual annotation by the Wikipedia users or using an automatic annotation process, allow us to determine the meaning and the saliency of a large number of words and phrases inside this corpus. These links are regarded as clues or salient features within the text that help define and disambiguate its context. We can measure the semantic relatedness of words by using their concept-based profiles, where a profile is constructed using the co-occurring salient concepts found within a given window size in a very large corpus.

To illustrate, let us consider the following paragraph extracted from a Wikipedia article:

> *An automobile, motor car or car is a wheeled motor vehicle used for transporting passengers, which also carries its own engine or motor. Most definitions of the term specify that automobiles are designed to run primarily on roads, to have seating for one to eight people, to typically have four wheels, and to be constructed principally for the transport of people rather than goods.*

All the underlined words and phrases represent linked concepts, which are disambiguated and connected to the correct Wikipedia article. We can therefore semantically interpret each term in this example as a vector of its neighboring linked concepts (as opposed to simple words, as done in the other corpus-based measures). For example the word $motor$ can be represented as a weighted vector of the salient concepts $automobile$, $motor\ car$, car, $wheel$, $motor\ vehicle$, $transport$, and $passenger$.

In this interpretation, a word is defined by a set of concepts which share its context and are weighted by their pointwise mutual information.

Our method consists of two main steps. In the first step, starting with Wikipedia, we create a corpus where concepts and saliency are explicitly annotated. Next, we use this corpus to build concept-based word profiles, which are used to measure the semantic relatedness of words and texts.

Constructing a Corpus Annotated for Concepts and Saliency

We create a large annotated corpus from Wikipedia, by linking salient words and phrases to their corresponding articles.

First, we use the manual links as provided by the Wikipedia users. These links have two important properties that are relevant to our method. On one hand, they represent concepts that are salient for a given context, since according to the Wikipedia guidelines, only those words or phrases that are important to the understanding of a certain text should be linked. On the other hand, the links connect surface forms to Wikipedia articles, thereby disambiguating the corresponding words or phrases. For instance, even if the word car is ambiguous, a link connecting this word to the Wikipedia article $motor\ car$ will eventually indicate that the intended meaning is that of $automobile$ rather than $railcar$.

Next, we use the one sense per discourse heuristic (Gale, Church, and Yarowsky 1992), according to which several occurrences of the same word within a discourse tend to share the same meaning. In our case, each additional occurrence of a word or phrase that matches a previously seen linked concept inside a given page is also linked to the same Wikipedia article. Moreover, since the already linked concepts are assumed to be salient for the given text, this property is transferred to the newly linked words or phrases. For instance, the second occurrence of the word $automobile$ in the previous example is also disambiguated and linked to a Wikipedia article even though it was not initially linked by a Wikipedia user. Additionally, since the first occurrence of $automobile$ was considered to be salient for this particular text (because of the first link), we assume that the second occurrence will also have this property.

Finally, we use a disambiguation method similar to the one used in the Wikify! system (Mihalcea and Csomai 2007), which assigns Wikipedia articles to words or phrases that have a high hyperlinkability (or keyphraseness). Very briefly, this method first determines the phrases that have a high probability (≥ 0.5) to be selected as a keyphrase, which corresponds to a high saliency. This probability is calculated as the number of times a word or phrase appears inside a manually assigned link divided by the total number of times that word or phrase appears in Wikipedia (hyperlinked or not). From this set, the words or phrases that have a probability of 95% or higher to point to only one article are tagged with the corresponding article. This disambiguation

method can be interpreted as a strengthened most frequent sense heuristic.

Overall, for a sample of 1 million articles, a total of ≈ 80 million salient concepts were collected. ≈ 40 million (51%) of these were manually disambiguated by the Wikipedia contributors, $\approx$ 30 million (33%) were disambiguated using the one sense per discourse heuristic, and ≈ 10 million (16%) were disambiguated with the strengthened most frequent sense heuristic.

Word Relatedness

We process our corpus to generate semantic profiles for words using their most contextually relevant concepts, which consist of the concepts linked to Wikipedia articles. To calculate the semantic relatedness of a given word pair, the overlap between the semantic profiles of the words in the word-pair is aggregated to produce a relatedness score.

Formally, given a corpus C with m tokens, vocabulary size N, and concept size W (number of unique Wikipedia concepts), a co-occurrence $N \times W$ matrix (E) is generated representing the accumulative co-occurrence frequencies of each of the corpus terms with respect to its context concepts (defined by a context window of size k). The elements of E are defined as follows:

$$E_{ij} = f^k(w_i, c_j) \tag{1}$$

where f^k is the number of times the terms w_i and concept c_j co-occur together within a window of k words in the entire corpus. The matrix is further processed to generate an $N \times W$ PMI matrix P, with elements defined as:

$$P_{ij} = log_2 \frac{f^k(w_i, c_j) \times m}{f^C(w_i) \times f^C(c_j)} \tag{2}$$

where $f^C(w_i)$ and $f^C(c_j)$ are the corpus frequencies for the term w_i and concept c_j respectively.

Each row P_i is further filtered to eliminate irrelevant associations by only keeping the top β_i cells (Islam and Inkpen 2006) and zeroing the rest. This corresponds to selecting the β highest scoring PMI terms associated with a given row:

$$\beta_i = (log_{10}(f^C(w_i)))^2 \times \frac{log_2(N)}{\delta}, \delta \geq 1 \tag{3}$$

where δ is a constant that is adjusted based on the size of the chosen corpus. To calculate the semantic relatedness between two words given the constructed matrix, we adopt a modified cosine-metric illustrated in equation 4.

$$Score_{cos}(A, B) = \frac{\sum_{y=1}^{N}(P_{iy} * P_{jy})^\gamma}{\sqrt{\sum_{y=1}^{N} P_{iy}^{2\gamma} * \sum_{y=1}^{N} P_{jy}^{2\gamma}}}, \tag{4}$$

The γ parameter allows us to control the weight bias. Additionally, since cosine is a normalized metric that scores one for identical terms, it is negatively impacted by a sparse space as it tends to provide low scores for near synonyms. This creates a large semantic gap between matching terms and strongly related terms. To close this gap and provide more meaningful scores, we also include a normalization factor λ, as shown in equation 5.

$$Sim(A, B) = \begin{cases} 1 & Score_{cos}(A, B) > \lambda \\ Score_{cos}(A, B)/\lambda & Score_{cos}(A, B) \leq \lambda \end{cases} \tag{5}$$

To further evaluate our model, we also adopt a slightly modified version of the Second Order Co-Occurrence Pointwise Mutual Information (SOCPMI), previously introduced in (Islam and Inkpen 2006). The measure was demonstrated to be a stronger metric compared to the traditional cosine similarity. According to the modified SOCPMI, the semantic association of two words A and B, with the corresponding rows P_i and P_j, is calculated as follows:

$$Score_{soc}(A, B) = ln(\frac{(\sum_{y=1}^{N}(P_{iy})^\gamma)}{\beta_i} + \frac{(\sum_{y=1}^{N}(P_{jy})^\gamma)}{\beta_j} + 1), \tag{6}$$

where $P_{iy} > 0, P_{jy} > 0$, and γ is a constant that controls the degree of bias toward terms with high PMI values. Since the resulted scores are not normalized, a normalization factor λ is also used in the same way as shown in equation 5, but using $Score_{soc}$ instead of $Score_{cos}$.

For the remainder of the paper we will refer to the system evaluated using SOCPMI metric over the concept space as SSA_s and the system evaluated using cosine as SSA_c. Mentions of SSA will address both metrics.

Text-to-Text Relatedness

To calculate the semantic relatedness between two text fragments, we use the same word profiles built from salient encyclopedic concepts, coupled with a simplified version of the text-to-text relatedness technique proposed in (Mihalcea, Corley, and Strapparava 2006) and (Islam and Inkpen 2009).

Formally, let T_a and T_b be two text fragments of size a and b respectively. After removing all stopwords, we first determine the number of shared terms (ω) between T_a and T_b. Second, we calculate the semantic relatedness of all possible pairings between non-shared terms in T_a and T_b. We further filter these possible combinations by creating a list φ which holds the strongest semantic pairings between the fragments' terms, such that each term can only belong to one and only one pair.

$$Sim(T_a, T_b) = \frac{(\omega + \sum_{i=1}^{|\varphi|} \varphi_i) \times (2ab)}{a + b} \tag{7}$$

where ω is the number of shared terms between the text fragments and φ_i is the similarity score for the ith pairing.

Experiments and Evaluations

When it comes to evaluating relatedness measures, the literature is split on the correct correlation to use. While a number of previous projects adopted the Pearson correlation metric r (Jarmasz 2003; Mohler and Mihalcea 2009; Islam and Inkpen 2006), there are several others that employed the Spearman correlation ρ (Gabrilovich and Markovitch 2007; Zesch, Muller, and Gurevych 2008).

We believe both metrics are important for the evaluation of semantic relatedness, where a good system should maintain the correct ranking between word pairs, and at the same time correctly quantify the strength of the relatedness for a given word-pair. We are therefore reporting both correlation metrics, as well as the harmonic mean of the Pearson and Spearman metrics $\mu = \frac{2r\rho}{r+\rho}$, which evaluates the ability of a system to simultaneously achieve the two goals of correct ranking and correct quantification.

Word Relatedness

To evaluate the effectiveness of the SSA model on word-to-word relatedness, we use three standard datasets that have been widely used in the past:

Rubenstein and Goodenough consists of 65 word pairs ranging from synonymy pairs (e.g., car - $automobile$) to completely unrelated terms (e.g., $noon$ - $string$). The 65 noun pairs were annotated by 51 human subjects. All the nouns pairs are non-technical words scored using a scale from 0 (not-related) to 4 (perfect synonymy).

Miller-Charles is a subset of the Rubenstein and Goodenough dataset, consisting of 30 word pairs. The relatedness of each word pair was rated by 38 human subjects, using a scale from 0 to 4.

WordSimilarity-353, also known as Finkelstein-353, consists of 353 word pairs annotated by 13 human experts, on a scale from 0 (unrelated) to 10 (very closely related or identical). The Miller-Charles set is a subset in the WordSimilarity-353 data set. Unlike the Miller-Charles data set, which consists only of single generic words, the WordSimilarity-353 set also includes phrases (e.g., *"Wednesday news"*), proper names and technical terms, therefore posing an additional degree of difficulty for any relatedness metric.

Parameter Tuning To choose the values for the δ, λ, and γ parameters for both SSA_s and SSA_c, we construct two additional tuning datasets, namely $HM30$ and $HM65$. The datasets are created by replacing $MC30$ and $RG65$ words with synonyms (e.g., replace *lad* with *chap*) or replacing the word-pair with a semantically parallel pair (e.g., replace *bird-crane* with *animal-puma*). Hence, the datasets are similar in terms of word relations they cover, yet they use completely different words. The parameters were adjusted to maximize the correlation on the two tuning datasets. The best matching set of parameters across the two datasets are $\delta = 0.3$, $\lambda = 0.125$, and $\gamma = 1$ for SSA_s, and $\delta = 0.4$, $\lambda = 0.01$, and $\gamma = 0.05$ for SSA_c, and these are the values used in all our experiments.

Results Table 1 shows the results obtained using our SSA relatedness model, compared to several state-of-the-art systems: knowledge-based methods including Roget and WordNet Edges (WNE) (Jarmasz 2003), $H\&S$ (Hirst and St Onge 1998), $J\&C$ (Jiang and Conrath 1997), $L\&C$ (Leacock and Chodorow 1998), Lin (Lin 1998), $Resnik$ (Resnik 1995); and corpus-based measures such as ESA (as published in (Gabrilovich and Markovitch 2007) and as obtained using our own implementation[3]), LSA (Landauer et al. 1997), and $SOCPMI$ (Islam and Inkpen 2006). Excluding LSA, ESA_{ours}, SSA_s, and SSA_c, which were implemented by us, the other reported results are based on the collected raw data from the respective authors. Some raw data was publicly available in previous publications (Li et al. 2006; Jarmasz 2003), otherwise we obtained it directly from the authors. Using the raw data, we recalculated the reported scores using the chosen metrics. The table also shows the weighted average WA for the three data sets, with the correlation weighted by the size of each dataset.[4]

The first examination of the results shows that the knowledge-based methods give very good results for the $MC30$ and $RG65$ datasets, which is probably explained by the deliberate inclusion of familiar and frequently used dictionary words in these sets. This performance quickly degrades on the $WS353$ dataset, largely due to their low coverage: the $WS353$ dataset includes proper nouns, technical and culturally biased terms, which are not covered by a typical lexical resource. This factor gives an advantage to the corpus-based measures like LSA and ESA, therefore achieving the best Spearman results on the $WS353$ dataset.

SSA consistently provides the best scores reporting a Pearson (r) weighted average of $0.649 - 0.671$ and Spearman (ρ) weighted average of $0.653 - 0.670$. The reported harmonic mean of Pearson and Spearman ($\mu = 0.651 - 0.671$) summarizes the performance of the SSA and ranks it as the best across all the datasets with an error reduction of $15.3\% - 20\%$ with respect to the closest baseline (LSA), surpassing even the knowledge-based methods.

It is also interesting to note that the SSA_s performance is superior to the $SOCPMI$ system despite the fact that SSA_s uses a similar metric. This implies that the disambiguated salient encyclopedic features used in SSA contribute significantly to the SSA_s's superior performance, much like ESA's superior performance being due to its use of the Wikipedia concept-space as compared to a typical vector-space model. A similar behavior is also evident in the SSA_c scores.

Text Relatedness

To evaluate the SSA model on text-to-text relatedness, we use three datasets that have been used in the past:

Lee50 (Lee, Pincombe, and Welsh 2005) is a compilation of 50 documents collected from the Australian Broadcasting Corporation's news mail service. Each document is scored by ten annotators based on its semantic relatedness to all the

[3]Since the published ESA results are limited to $MC30$, $WS353$, and $LEE50$, we resolved to use our own ESA implementation to cover the rest of the datasets for a more meaningful comparison. It is worth noting that our implementation provides better Pearson score (0.744) for $MC30$ than the one reported by (Gabrilovich and Markovitch 2007) (0.588) while managing to provide equivalent Pearson scores for $WS353$. Additionally, it outperforms other ESA implementations (Zesch, Muller, and Gurevych 2008).

[4]Throughout this paper, best results in each column are formatted in bold, while second best results are underlined.

	r				ρ				μ			
Metric	**MC30**	**RG65**	**WS353**	**WA**	**MC30**	**RG65**	**WS353**	**WA**	**MC30**	**RG65**	**WS353**	**WA**
$Roget$	0.878	0.818	0.536	0.600	**0.856**	0.804	0.415	0.501	**0.867**	0.814	0.468	0.545
WNE	0.732	0.787	0.271	0.377	0.768	0.801	0.305	0.408	0.749	0.796	0.287	0.392
$H\&S$	0.689	0.732	0.341	0.421	0.811	0.813	0.348	0.446	0.745	0.772	0.344	0.433
$J\&C$	0.695	0.731	0.354	0.432	0.820	0.804	0.318	0.422	0.753	0.767	0.335	0.426
$L\&C$	0.821	0.852	0.356	0.459	0.768	0.797	0.302	0.405	0.793	0.828	0.327	0.431
Lin	0.823	0.834	0.357	0.457	0.750	0.788	0.348	0.439	0.785	0.810	0.352	0.447
$Resnik$	0.775	0.800	0.365	0.456	0.693	0.731	0.353	0.431	0.732	0.770	0.359	0.444
ESA_{Gab}	0.588	$--$	0.503	$--$	0.727	$--$	**0.748**	$--$	0.650	$--$	0.602	$--$
ESA_{our}	0.744	0.716	0.492	0.541	0.704	0.749	0.435	0.499	0.723	0.732	0.461	0.518
LSA	0.725	0.644	0.563	0.586	0.662	0.609	0.581	0.590	0.692	0.626	0.572	0.588
SOC_{PMI}	0.764	0.729	$--$	$--$	0.78	0.741	$--$	$--$	0.772	0.735	$--$	$--$
SSA_s	0.871	0.847	**0.622**	**0.671**	0.810	0.830	0.629	**0.670**	0.839	0.838	**0.626**	**0.671**
SSA_c	**0.879**	**0.861**	0.590	0.649	0.843	**0.833**	0.604	0.653	0.861	**0.847**	0.597	0.651

Table 1: Pearson (r), Spearman (ρ) and their harmonic mean (μ) correlations on the word relatedness datasets. The weighted average WA over the three datasets is also reported.

	r				ρ				μ			
Metric	**Li30**	**Lee50**	**AG400**	**WA**	**Li30**	**Lee50**	**AG400**	**WA**	**Li30**	**Lee50**	**AG400**	**WA**
ESA_{our}	0.810	0.635[2]	0.425	0.584	0.812	0.437	0.389	0.434	0.811	0.518	0.406	0.498
LSA	0.838	**0.696**	0.365	0.622	0.863	0.463	0.318	0.433	0.851	0.556	0.340	0.512
Li	0.81	$--$	$--$	$--$	0.801	$--$	$--$	$--$	0.804	$--$	$--$	$--$
STS	0.848	$--$	$--$	$--$	0.832	$--$	$--$	$--$	0.840	$--$	$--$	$--$
SSA_s	**0.881**	0.684	**0.567**	**0.660**	0.878	0.480	**0.495**	0.491	**0.880**	0.564	**0.529**	0.561
SSA_c	0.868	0.684	0.559	0.658	0.870	**0.488**	0.478	**0.492**	0.869	**0.569**	0.515	**0.562**

Table 2: Pearson (r), Spearman (ρ) and their harmonic mean (μ) correlations on the text relatedness datasets. The weighted average WA over the three datasets is also reported.

other documents. The users' annotation is then averaged per document pair, resulting in 2,500 document pairs annotated with their similarity scores. Since it was found that there was no significant difference between annotations given a different order of the documents in a pair (Lee, Pincombe, and Welsh 2005), the evaluations are carried out on only 1225 document pairs after ignoring duplicates.

Li30 (Li et al. 2006) is a sentence pair similarity dataset obtained by replacing each of the Rubenstein and Goodenough word-pairs (Rubenstein and Goodenough 1965) with their respective definitions extracted from the Collins Cobuild dictionary (Sinclair 2001). Each sentence pair was scored by 32 native English speakers, and the scores were then averaged to provide a single relatedness score per sentence-pair. Due to the resulted skew in the scores toward low similarity sentence-pairs, a subset of 30 sentences was manually selected from the 65 sentence pairs to maintain an even distribution across the similarity range (Li et al. 2006).

AG400 (Mohler and Mihalcea 2009) is a domain specific dataset from computer science, used to evaluate the application of semantic relatedness measures to real world applications such as short answer grading. The original dataset consists of 630 student answers along with the corresponding questions and correct instructor answers. Each student answer was graded by two judges on a scale from 0 to 5, where 0 means completely wrong and 5 represents a perfect answer. The Pearson correlation between human judges was measured at 0.64. Since we noticed a large skew in the grade distribution toward the high end of the grading scale (over 45% of the answers are scored 5 out of 5), we followed (Li et al. 2006) and randomly eliminated 230 of the highest grade answers in order to produce more normally distributed scores and hence calculate a meaningful Pearson correlation.

Parameter Tuning Instead of creating a tuning dataset for the text-to-text relatedness task, we opted to rely on the parameters selected for the word-to-word relatedness task.

Results Table 2 shows the text relatedness results for the $Li30$, $Lee50$, and $AG400$ datasets. The results are compared with several state-of-the-art systems: ESA (Gabrilovich and Markovitch 2007), LSA (Landauer et al. 1997), and STS (Islam and Inkpen 2008). As seen in Table 2, SSA_c and SSA_s are clear winners, even when compared to the STS system, which relies on the SOCPMI framework. While LSA provides the best Pearson score for $Lee50$ ($r = 0.696$), its superiority does not extend to Spearman ($\rho = 0.463$). This inconsistency is penalized as seen in the harmonic mean score ($\mu = 0.512$). It is also interesting to see the large improvements achieved by SSA_s ($\mu = 0.529$) and SSA_c ($\mu = 0.515$) over the LSA ($\mu = 0.340$) and ESA ($\mu = 0.406$) when evaluated on the $AG400$ dataset. To explore this in more detail, and also for a comparison with other knowledge-based and corpus-based measures, Table 3 shows a comparison of the SSA_s and SSA_c systems with all other relatedness measures, as reported by (Mohler and Mihalcea 2009). As it was the case in the word relatedness evaluations, SSA_s and SSA_c display a performance that is superior to all the knowledge-based and corpus-based metrics, with an error-reduction of $10.6\% - 13.3\%$ in harmonic mean with respect to the closest competitor ($J\&C$).

Conclusions

In this paper, we proposed a novel unsupervised method for semantic relatedness that generates a semantic profile for words by using salient conceptual features gathered from encyclopedic knowledge. The model is built on the notion that the meaning of a word can be represented by the

Measure	r	ρ	μ
Knowledge-based measures			
WNE	0.440	0.408	0.424
$L\&C$	0.360	0.152	0.214
$Lesk$	0.382	0.346	0.363
$Wu\&Palmer$	0.456	0.354	0.399
$Resnik$	0.216	0.156	0.181
Lin	0.402	0.374	0.388
$J\&C$	0.480	0.436	0.457
$H\&S$	0.243	0.192	0.214
Corpus-based measures			
LSA	0.365	0.318	0.340
ESA_{ours}	0.425	0.389	0.406
SSA_s	**0.567**	**0.495**	**0.529**
SSA_c	0.559	0.478	0.515
Baseline			
$tf * idf$	0.369	0.386	0.377

Table 3: Comparative results using Pearson (r), Spearman (ρ) and their harmonic mean (μ) for the AG400 dataset, for the metrics reported in (Mohler and Mihalcea 2009)

salient concepts found in its immediate context. The evaluation on standard word-to-word and text-to-text relatedness benchmarks confirms the superiority and consistency of our model. The performance of the model seems to be independent of the distance metric used in our evaluation ($cosine$ or $SOCPMI$). This fact provides additional support for the underlying assumption about profiling words using strong unambiguous word-concept associations.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation CAREER award #0747340 and IIS award #1018613. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Church, K. W., and Hanks, P. 1990. Word association norms, mutual information, and lexicography. *Comput. Linguist.* 16(1):22–29.

Fellbaum, C., ed. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Gabrilovich, E., and Markovitch, S. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1606–1611.

Gale, W.; Church, K.; and Yarowsky, D. 1992. One sense per discourse. In *Proceedings of the DARPA Speech and Natural Language Workshop*, 233–237.

Hirst, G., and St Onge, D. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In Fellbaum, C., ed., *WordNet: An Electronic Lexical Database*, 305–332. MIT Press.

Islam, A., and Inkpen, D. 2006. Second order co-occurrence pmi for determining the semantic similarity of words. *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2006)* 1033–1038.

Islam, A., and Inkpen, D. 2008. Semantic text similarity using corpus-based word similarity and string similarity. *ACM Trans. Knowl. Discov. Data* 2(2):1–25.

Islam, A., and Inkpen, D. 2009. Semantic Similarity of Short Texts. In Nicolov, N.; Angelova, G.; and Mitkov, R., eds., *Recent Advances in Natural Language Processing V*, volume 309 of *Current Issues in Linguistic Theory*. Amsterdam & Philadelphia: John Benjamins. 227–236.

Jarmasz, M. 2003. *Roget's thesaurus as a Lexical Resource for Natural Language Processing*. Ph.D. Dissertation, Ottawa-Carleton Institute for Computer Science, School of Information Technology and Engineering, University of Ottawa.

Jiang, J. J., and Conrath, D. W. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *International Conference Research on Computational Linguistics (ROCLING X)*, 9008+.

Landauer, T. K.; L, T. K.; Laham, D.; Rehder, B.; and Schreiner, M. E. 1997. How well can passage meaning be derived without using word order? a comparison of latent semantic analysis and humans.

Leacock, C., and Chodorow, M. 1998. *Combining Local Context and WordNet Similarity for Word Sense Identification*. The MIT Press. chapter 11, 265–283.

Lee, M. D.; Pincombe, B.; and Welsh, M. 2005. An empirical evaluation of models of text document similarity. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, 1254–1259. Mahwah, NJ: Erlbaum.

Lesk, M. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *SIGDOC '86: Proceedings of the 5th annual international conference on Systems documentation*, 24–26.

Li, Y.; McLean, D.; Bandar, Z. A.; O'Shea, J. D.; and Crockett, K. 2006. Sentence similarity based on semantic nets and corpus statistics. *IEEE Trans. on Knowl. and Data Eng.* 18(8):1138–1150.

Lin, D. 1998. An information-theoretic definition of similarity. In *ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning*, 296–304.

Mihalcea, R., and Csomai, A. 2007. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management*.

Mihalcea, R.; Corley, C.; and Strapparava, C. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the Association for Artificial Intelligence*, 775–780.

Mohler, M., and Mihalcea, R. 2009. Text-to-text semantic similarity for automatic short answer grading. In *Proceedings of the European Association for Computer Linguistics*, 567–575.

Resnik, P. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 448–453.

Rubenstein, H., and Goodenough, J. B. 1965. Contextual correlates of synonymy. *Commun. ACM* 8(10):627–633.

Sinclair, J. 2001. *Collins Cobuild English Dictionary for Advanced Learners*, volume third edition. Harper Collins.

Turney, P. 2001. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)*.

Wu, Z., and Palmer, M. 1994. Verb semantics and lexical selection. In *Proceedings of the Association for Computational Linguistics*, 133 –138.

Zesch, T.; Muller, C.; and Gurevych, I. 2008. Using wiktionary for computing semantic relatedness. In *Proceedings of AAAI*, 861–867.