Journal of Intelligent Information Systems manuscript No.
(will be inserted by the editor)

Sentence Similarity based on Semantic Kernels for Intelligent Text Retrieval

Samir Amir, Adrian Tanasescu and Djamel A. Zighed

Received: date / Accepted: date

Abstract We propose a new approach to compute semantic similarity between sentences. It is based on the semantic kernel, composed of subject, verb, and object that, we suppose, summarize the general meaning of each sentence. Thanks to linguistics resources available such as Stanford Parser, many features are then extracted from the semantic kernels and aggregated by mean of weights. The weighting is produced by a supervised machine learning technique on a training data set provided by human experts as ground truth. The cross validation shows good performances. Thanks to this similarity measure between sentences, one can build an intelligent text retrieval engine more sensitive to the semantic content of texts, specifically suited for short texts than the classical methods based on bag of words. An application is being developed for highlighting parts of speech in scientific articles.

1 Introduction

With the growing of Internet and the increasing of its usability, the ability to determine the semantic similarity between texts has proven useful for a wide variety of applications. Question answering [1], recommendation systems [23], and conversational agents [12] are examples where computers have to understand human language and hidden meanings. The main objective of text similarity approaches is to improve the ability of computers in comprehending human language which may use several ways to express the same meaning.

Measuring the similarity between sentences using traditional Information Retrieval (IR) [21] or Natural Language Processing (NLP) [12] approaches is not always efficient. This is due to the fact the sentences may not contain any common words or the co-occurrence of words is rarely present. In this context,

Institut des Sciences de l'Homme,
14 Avenue Berthelot, 69007 Lyon, France.
{samir.amir, adrian.tanasescu, abdlkader.zighed}@ish-lyon.cnrs.fr

various short text similarity measures have been recently proposed. However, there is still place for improvement to do due to the complexity of human language.

Sentences may have an equivalent meaning with heterogeneous structures in terms of word positions, number of words as well as the grammatical relations. For instance, according to the benchmarks described in [18], the two sentences "*I am so hungry I could eat a whole horse plus dessert.*" and "*I could have eaten another meal, I'm still starving.*" have been scored at 0.765 by human experts while we can clearly see that their structures are heterogeneous in terms of number of words and word positions.

Our research is oriented towards finding an efficient way of detecting paragraphs in texts that correspond to a given phrase. For instance, we wish to be able to detect paragraphs in research articles that develop the ideas expressed in the phrases of their abstract. The potential use of such an approach is tremendous in terms of semantic text retrieval. The existing approaches for text similarity are either still theoretical or not available for reuse. Furthermore, scalability issues are rarely addressed. In order to deal with these issues, we focused on proposing a new sentence similarity measure based on semantic kernels. We used Stanford Parser [5] as an NLP tool to parse sentences. Then, based on the grammatical relationships returned by the latter, we constructed the kernels which are triples composed of subjects, verbs, and objects. We consider that kernels reflect the general meaning of sentences. This offers a certain flexibility to deal with heterogeneous structures and allows us to accelerate computations of text similarities. Based on extracted kernels, an analysis is conducted to extract the most relevant information (features). To do so, a machine learning technique is used to find the best way for aggregating features in the a semantic model. The proposed method uses WordNet [10] and DBpedia [14] as semantic resources. Evaluation of the measure shows that our approach achieves better results then several reputed and recent methods in the literature.

The next section discusses some related work; Section 3 describes the proposed method for measuring sentence similarity. In Section 4, we show and discuss the experimentation evaluation for our approach. We conclude our work in Section 5

2 Related Work

With the emergence of Internet, several methods for computing similarities between texts have been proposed. Most of these studies are mostly dedicated for comparing documents and they are not optimized for processing short texts. In this section, we will discuss some of the methods reported in the state of the art described in [7].

Latent Semantic Analysis (LSA) is one of the most used in the literature [13], mainly for in the area of information retrieval. LSA is built on top of a statistical computing using the co-occurrence probability of words in sentences;

sentence similarities are computed using singular value decomposition (SVD). Vector space model (VSM) is another approach based on an algebraic model for representing text documents as vectors of features, such as, for example, index terms [22]. It is used in information filtering, information retrieval, indexing and relevancy rankings. A comparative study between both approaches is given in [20].

Among recent works concerning short text similarity computation we can cite STASIS (Sentence Similarity based on Semantic Nets and Corpus Statistics) [15]. The latter is an hybrid approach which uses WordNet [10] as a lexical database. Additionally, words are organized according to their meanings. STASIS uses VSM to build the sematic space and compute the similarity between sentences. SyMSS [17] is another method based on the notion that the meaning of a sentence is made up of not only the meanings of its individual words, but also the structure word combinations. SyMSS uses a lexical database to get more semantics about words and calculates the similarity between concepts playing the same syntactic role in the sentence. To do so, a psychological plausibility by weighting the different syntactic role is also introduced.

Omiotis [24] is a new similarity method constructed from the word thesaurus and lexical ontology WordNet [10]. The method is capable of handling the synonymy and polysemy problems. Omiotis exploits all of the semantic information a thesaurus can offer, including semantic relations crossing parts of speech (POS). TF-IDF method [21] is also used by Omiotis for term weighting using the text corpus.

LSS [4] is another approach which uses WordNet to determine the word semantic similarity. However, LSS does not consider the order of the words in the documents. LSS is able to identify highly similar meanings without respect to details.

STS [11] based on a modified version of the Longest Common Sub-sequence (LCS) string matching algorithm, is currently the most accurate algorithm in terms of proximity to human similarity. It uses set of metrics as word-string similarity and the longest common subsequence algorithm for computing the semantic similarity between sentences.

3 Semantic Kernels Based Approach

In order to assess similarity between phrases/sentences, we focus on kernels of sentences in terms of semantics. These kernels are given by subjects, verbs and objects. It is obvious that the meaning of a sentence resides in all of it's components. However we can also easily presume that the essence of its meaning can be found in what we call "the sentence kernel", namely its subject, verb and object [8]. This assumption allows us to concentrate on the essential part of a sentence and thus to considerably accelerate similarity computations. Note that sentences may be composed of several phrases. In this case, calculating semantic similarity between sentences needs to deal with comparisons

of kernels composing the sentences. Therefore, the approach we proposed also deals with computing overall similarity between multi-phrase sentences.

3.1 Preliminaries

Hereafter we define the concepts that we will use in the remaining of our paper.

A *sentence* is a text composed of one or more phrases separated by ponctuation or by conjunctions. A *phrase* is an assertion or an interrogation composed of one or several subjects, one verb and one or several objects. A phrase is composed of one or several *kernels*, each kernel being defined as a triple containing one subject, one verb and one object.

3.2 Kernels extraction

First, we use Stanford Parser to extract kernels from phrases [5]. The latter contains more than 50 grammatical relations to describe contents. Indeed, for each grammatical relation, we have defined a specific rule to extract kernels. For instance, for the sentence *Bell, based in Los Angeles, makes and distributes electronic, computer and building products*, Stanford Parser returns the following relations. Algorithm 1 gives an example of some rules to deal with the *nsubj* relation.

nsubj(makes-8, Bell-1)
nsubj(distributes-10, Bell-1)
vmod(Bell-1, based-3)
nn(Angeles-6, Los-5)
prep in(based-3, Angeles-6)
root(ROOT-0, makes-8)
conj and(makes-8, distributes-10)
amod(products-16, electronic-11)
conj and(electronic-11, computer-13)
amod(products-16, computer-13)
conj and(electronic-11, building-15)
amod(products-16, building-15)
dobj(makes-8, products-16)
dobj(distributes-10, products-16)

Table 1 An example of grammatical relations returned by Stanford Parser

Using the above-mentioned relations, we will define a more generic representation of a sentence. The latter will be presented as a set of kernels which represent the core of the composing phrases: (*Bell, make, electronic*), (*Bell, make, computer*), (*Bell, make, building products*), (*Bell, distribute, electronic*), (*Bell, distribute, computer*), and (*Bell, distributes, building products*). Using this representation will certainly cause a loss of information. However, according to some studies [8], the phrase kernel reflects the general meaning of

Algorithm 1 Example of rule-based approach for building kernels

```
1: Input parameters: Grs (grammatical relations returned by Stanford Parser)
2: i=0, j=0;
3: For all Gr in Grs do
4: if (Gr(i) = nsubj) then
5:     Create new kernel k(j)
6:     k(j).setSubjectName(Gr(i).getName())
7:     k(j).setVerb(Gr(i).getVerb())
8:     if (Gr(i-1) = nsubj) then
9:         j++
10:        Create new kernel k(j)
11:        k(j).setSubjectName(k(j-1).getName())
12:        k(j).setVerb(Gr(j).getVerb())
13:    end if
14:    if (Gr(i) = dobj) then
15:        // set the same object for the two previou kernels
16:        k(j).setObject(Gr(i).getName())
17:        k(j-1).setObject(Gr(i).getName())
18:        ...
19:    end if
20: end if
```

phrases. We have based our approach on this representation of sentences in sets of semantic kernels, which we estimate as an accurate enough approximation of their overall meaning. We will show in section 4 that the accuracy of this kernel representation of phrase semantics easily sustains the comparison with existing approaches.

3.3 Semantic similarity among kernel elements

This phase deals with the semantic similarity computation between kernel elements, namely subjects, verbs, and objects. After a lemmatization step, we apply Hirst and St-Onge Measure (HSO) [9] to calculate the similarity between kernel elements, using WordNet as linguistic resource. HSO calculates relatedness between terms using the path distance between their relative positions in the ontology schema. Our choice of using HSO among all existing measures was based on the evaluation study carried out in [3]. WordNet may not cover all the terms extracted from the sentences, in this case, we use DBpedia as a second resource to retrieve the type of term.

Finally, if both resources do not cover the term, Jaro-Winkler [25] distance is used as a string based measure. Algorithm 2 shows the different steps for computing the semantic similarity between two kernels.

Once the similarities between elements of two kernels are computed, we repeat this operation for all pairs of kernels composing the two sentences that are compared. This allows us to obtain a set of similarities between every subjects, objects and verbs composing the sentences.

Algorithm 2 Assess semantic similarities between two kernels

```
Input parameters: two kernels K1 and K2, from different sentences
Each kernel is a triple (subject, verb, object).
calculate HSO between subjects, verbs and objects.
if WordNet does not cover the the queried term. then
    query DBpedia
    if DBpedia does not cover the the queried term. then
        use Jaro-Winkler Distance
    end if
end if
return the similarities σs, σv, σo between subjects, verbs and objects belonging to K1
and K2
```

Let's consider two sentences P_1 and P_2. Let's suppose P_1 is composed of two kernels $K_1 = (s_1, v_1, o_1)$ and $K_2 = (s_2, v_2, o_2)$ while P_2 is composed of one kernel $K_3 = (s_3, v_3, o_3)$.

The expected preliminary output will be the following set of similarities between kernels of P_1 and P_2:

$\sigma(s_1, s_3), \sigma(s_2, s_3)$ - similarities between subjects

$\sigma(v_1, v_3), \sigma(v_2, v_3)$ - similarities between verbs

$\sigma(o_1, o_3), \sigma(o_2, o_3)$ - similarities between objects

In the approach that we present in section 3.4 we focus on studying a way of combining these partial similarities into an overall similarity score between two phrases.

Not that some existing approaches simply make a linear combination (sum) of similarities between subjects, verbs and objects in order to obtain overall similarities between sentences [16]. To the best our knowledge, the best way of estimating the importance of each element is to learn from ground truth. Therefore, in the next section we explain how to aggregate kernel element similarities into an overall phrase similarity measure.

3.4 Semantic Model

The semantic similarity model consists in the definition of a way of combining semantic comparisons of sentence kernel elements (subject, verb, object) in order to estimate overall similarity between two sentences. We propose to use machine learning in order to detect the best way to combine semantic similarities between sentence parts obtained from Algorithm 2. Therefore, we use the benchmark described in [19] containing couples of sentences already rated by humans in terms of perceived similarity.

3.4.1 Model definition

Previously, we have computed similarities between all kernels composing the sentences for which we wish to achieve the overall similarity. Hereafter, we

propose a way to combine these similarities into an overall similarity between sentences.

As stated before, considering two given sentences P_1 and P_2, a set of similarities are computed among subjects, verbs and objects composing the kernels in P_1 and P_2. For each kernel i in P_1 and each kernel j in P_2:

$X_{ij} = \sigma(s_i, s_j)$ the similarity between subjects of kernel i and j.
$Y_{ij} = \sigma(v_i, v_j)$ the similarity between verbs, and
$Z_{ij} = \sigma(o_i, o_j)$ the similarity between objects.

Using X_{ij}, Y_{ij} and Z_{ij}, we compute the following aggregated measures as features assessing the similarity between sentences:

- $\overline{X_{ij}}$ - the average of similarities X_{ij}
- $\overline{Y_{ij}}$ - the average of similarities Y_{ij}
- $\overline{Z_{ij}}$ - the average of similarities Z_{ij}
- $max(X_{ij})$ - the maximum X_{ij}
- $max(Y_{ij})$ - the maximum Y_{ij}
- $max(Z_{ij})$ - the maximum Z_{ij}
- N_{ij} - is the number of comparisons between kernels representing the two phrases being compared.

The main idea here is to determine, algorithmically, which similarity elements will be used and how they will be combined into an overall similarity measure between the compared phrases. For instance, intuitively we thought that similarity between verbs are less likely to play the same role in the overall similarity as the similarity between subjects. Therefore, and here relies one novelty of our approach, we propose to estimate the contributions of partial similarities considered herebeforein the overall similarity by using machine learning. We have chosen general linear modeling [6], a.k.a. linear regression, as the learning algorithm for its simplicity of interpretation and easyness of future deployment.

3.4.2 Model estimation

The goal of this step is to build an overall similarity measure that best estimates the human-assessed similarity scores between two sentences, based on partial aggregated measures related to similarity comparisons between kernels composing the sentences. In order to determine the best features to include in the overall similarity we have used forward stepwise linear regression [2]. Basically, this technique introduces, one step at a time, the best feature that will improve the overall estimation error.

The learning process was done on a subset of the benchmark composed of 30 couples of phrases. The human similarity to be estimated as well as the partial similarities considered were expressed as real numbers in the range [0,1]. The best model obtained using this approach included only three of the seven initial features. The overall similarity between two sentences will be described as:

$$\sigma(P_x, P_y) = -0.13213 + 0.60341 * \overline{X_{ij}} + 0.38057 * \overline{Z_{ij}} + 0.04893 * N_{ij}$$

We can observe that this model does not take into account any feature related to verbs similarity. This is due to the fact that these features often have high values (verbs are similar) while overall human similarity is very low. This does not mean that verbs similarity does not contribute to overall similarity but only that their contribution is not directly related to it.

The obtained model partially explains the variance of the human similarity that was estimated. Adjusted R^2, assessing the quality of the linear regression, was estimated at 0.66 while residual standard error (RSE) established at 0.16.

The evaluation of this model on the entire benchmark described in [19] revealed a correlation of 0.776 with the average human similarity and an average deviation of 0.126. In the following section we compare the results obtained by the model we proposed with other approaches that attempt to estimate similarities between couples of phrases.

4 Model Assessment

As stated before, we have used the benchmark described in [19] in the construction of our semantic model estimating the similarities. Further on, in order to assess the generalization capability of our model outside of the initial benchmark used for the model construction, we have also used a subset of sentences described in a second benchmark [18]. We first compare the results of our model to other existing approaches already proposing measures of similarity on the benchmark described in [19]. Then we discuss the way our approach generalizes on the second benchmark.

4.1 Comparisons based on the initial benchmark

Hereafter, we have compared the results of our approach, that we named *SK*, to five other approaches that have been already tested on a partial subset of the benchmark we used for training our model [4]. For each approach presented we have listed in Table 2 similarity values for each one of the 30 sentence pairs tested.

We have computed indicators of correlation between each approach and Human similarity as well as their average deviation from the human similarity. We can observe that our approach (SK) easily bears the comparison with existing approaches.

Correlation with human assessed similarity (column 2) establishes at 0.83414, just behind and very close to STS and Omiotis. However in terms of average deviation, SK comes second closely behind STS with much better results than LSA, LSS and STASIS as shown in Figure 1.

Note that in the context of our main objective of building a fast semantically similar text retrieval engine, the mean deviation is a more relevant

Table 2 Comparison of approaches on benchmark described in [19]

SP	Human	LSS	STASIS	LSA	Omiotis	STS	SK
1	0.010	0.180	0.329	0.510	0.1062	0.06	0.084
5	0.005	0.198	0.287	0.530	0.1048	0.11	0.084
9	0.005	0.280	0.209	0.505	0.1046	0.07	0.013
13	0.108	0.166	0.530	0.535	0.3028	0.16	0.406
17	0.063	0.324	0.356	0.575	0.2988	0.26	0.243
21	0.043	0.324	0.512	0.530	0.2430	0.16	0.169
25	0.065	0.220	0.546	0.595	0.2995	0.33	0.325
29	0.013	0.220	0.335	0.505	0.1074	0.12	0.159
33	0.145	0.324	0.590	0.810	0.4946	0.29	0.312
37	0.130	0.280	0.438	0.580	0.1085	0.20	0.314
41	0.283	0.324	0.428	0.575	0.1082	0.09	0.079
47	0.348	0.198	0.721	0.715	0.2164	0.30	0.185
48	0.355	1.000	0.641	0.615	0.5295	0.34	0.453
49	0.293	1.000	0.739	0.540	0.5701	0.15	0.277
50	0.470	0.800	0.685	0.675	0.5502	0.49	0.299
51	0.138	0.800	0.649	0.725	0.5206	0.28	0.275
52	0.485	1.000	0.493	0.695	0.5987	0.32	0.294
53	0.483	0.471	0.394	0.830	0.4965	0.44	0.446
54	0.360	0.800	0.517	0.610	0.4255	0.41	0.141
55	0.405	0.800	0.550	0.700	0.4287	0.19	0.253
56	0.588	0.800	0.759	0.780	0.9308	0.47	0.482
57	0.628	1.000	0.700	0.750	0.6120	0.26	0.65
58	0.590	0.800	0.753	0.830	0.7392	0.51	0.495
59	0.863	1.000	1.000	0.985	0.9982	0.94	0.900
60	0.580	0.800	0.663	0.830	0.9309	0.60	0.359
61	0.523	0.800	0.662	0.630	0.3466	0.29	0.343
62	0.773	1.000	0.729	0.740	0.7343	0.51	0.686
63	0.558	1.000	0.639	0.870	0.7889	0.52	0.662
64	0.955	1.000	0.998	1.000	0.9291	0.93	0.900
65	0.653	1.000	0.831	0.860	0.8194	0.65	0.589

Table 3 Comparison of approaches on benchmark described in [18]

SP	Human	LSA	SK	SP	Human	LSA	SK
66	0.252	0.45	0.416	102	0.272	0.28	0.374
69	0.025	0.25	0.027	105	0.192	0.03	0.325
74	0.180	0.37	0.278	108	0.752	0.23	0.900
75	0.362	0.06	0.215	111	0.040	0.05	0.129
77	0.032	0.23	0.226	112	0.282	0.49	0.267
80	0.125	0.15	0.041	114	0.900	0.72	0.953
84	0.512	0.58	0.416	115	0.220	0.39	0.137
85	0.427	0.64	0.520	117	0.137	0.07	0.267
86	0.020	0.01	0.064	118	0.300	0.3	0.121
92	0.580	0.53	0.447	119	0.170	0.06	0.386
93	0.425	0.49	0.255	122	0.030	0.07	0.038
94	0.490	0.160	0.274	125	0.237	0.02	0.116
98	0.055	0.140	0.011	128	0.125	0.05	0.199
99	0.990	1	0.900	130	0.040	0.0	0.003
100	0.262	0.58	0.178	131	0.765	0.54	0.666

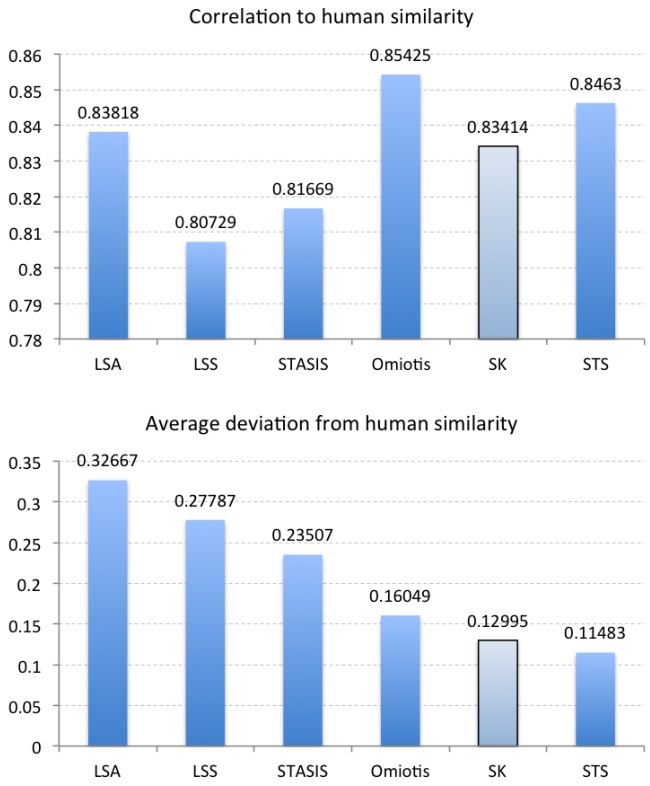

Fig. 1 Correlations and mean deviation of the compared approaches.

evaluation measure as the goal here is to obtain the closest estimation to human similarity.

In Table 3 we have presented the way our approach perform on a different benchmark and we have also reported results obtained by LSA (General Reading up to 1st year college - 300 factors)[1]. The correlations with human similarity establish at 0.73676 for LSA and 0.89548 for SK. Average deviations are 0.14405 and 0.10488 respectively. Note that LSA was the only approach for which we could provide comparison values as all other approaches were unavailable to our usage.

Our experiments show that similarity measure that we built, based on kernels that approximate sentences, easily bares the comparison with the existing approaches assessing semantic similarity.

[1] http://lsa.colorado.edu/

5 Conclusion and future work

In this paper we presented a new approach for assessing short text similarity. The great strength of our approach is that it is based on semantic kernels which offer certain flexibility in terms of similarity computation. Based on an existing benchmark, we used a learning technique to extract features from kernels and build the corresponding semantic model. The proposed approach was also strengthened by using several external resources and semantic measures. Experimental evaluation showed that our approach approximates human understanding at least as well as the existing approaches. Moreover, the results showed that the semantic model is robust. Indeed, a high correlation was obtained with a benchmark which is different from that used in the training process.

Although the obtained results are encouraging, we think that there is still place for improvement on all levels. Firstly, our approach relies on Stanford Parser. Thus, in order to fully use the potential of the latter, additional rules may be defined to consider other dimensions such as the context of the sentence. Secondly, we think that the proposed semantic model can be extended by adding more feature as temporal aspects. Finally, we believe that the training set we used in our experiment is limited and does not represent all possible cases we can find in real life. For this reason, we also simultaneously concentrate our efforts on building an alternate data-set that we will submit for human similarity assessment shortly.

6 Appendix

Table 4: the benchmark used for the first experiment [19]

ID	**Sentence**	**human assessed similarity**
1	Cord is strong, thick string. A smile is the expression that you have on your face when you are pleased or amused, or when you are being friendly.	0.010
5	An autograph is the signature of someone famous which is specially written for a fan to keep. The shores or shore of a sea, lake or wide river is the land along the edge of it.	0.005
9	An Asylum is a psychiatric hospital. Fruit or a fruit is something which grows on a tree or bush and which contains seeds or a stone covered by a substance that you can eat.	0.005
13	A boy is a child who will grow up to be a man.	0.108

	A rooster is an adult male chicken.	
17	The coast is an area of land that is next to the sea. A forest is a large area where trees grow close together.	0.063
21	A boy is a child who will grow up to be a man. A sage is a person who is regarded as being very wise.	0.043
25	A forest is a large area where trees grow close together. A graveyard is an area of land, sometimes near a church, where dead people are buried.	0.065
29	A bird is a creature with feathers and wings, females lay eggs and most birds can fly. Woodland is land with a lot of trees.	0.013
33	A hill is an area of land that is higher than the land that surrounds it. Woodland is land with a lot of trees.	0.145
37	A magician is a person who entertains people by doing magic tricks. In ancient times, an oracle was a priest or priestess who made statements about future events or about the truth.	0.130
41	In ancient times, an oracle was a priest or priestess who made statements about future events or about the truth. A sage is a person who is regarded as being very wise.	0.283
47	A furnace is a container or enclosed space in which a very hot fire is made, for example to melt metal, burn rubbish or produce steam. A stove is a piece of equipment which provides heat, either for cooking or for heating a room.	0.348
48	A magician is a person who entertains people by doing magic tricks. In legends and fairy stories, a wizard is a man who has magic powers.	0.355
49	A hill is an area of land that is higher than the land that surrounds it. A mound of something is a large rounded pile of it.	0.293
50	Cord is strong, thick string. String is thin rope made of twisted threads, used for tying things together or tying up parcels.	0.470

51	Glass is a hard transparent substance that is used to make things such as windows and bottles. A tumbler is a drinking glass with straight sides.	0.138
52	A grin is a broad smile. A smile is the expression that you have on your face when you are pleased or amused, or when you are being friendly.	0.485
53	In former times, serfs were a class of people who had to work on a particular person's land and could not leave without that person's permission. A slave is someone who is the property of another person and has to work for that person.	0.483
54	A When you make a journey, you travel from one place to another. A voyage is a long journey on a ship or in a spacecraft.	0.360
55	An autograph is the signature of someone famous which is specially written for a fan to keep. Your signature is your name, written in your own characteristic way, often at the end of a document to indicate that you wrote the document or that you agree with what it says.	0.405
56	The coast is an area of land that is next to the sea. The shores or shore of a sea, lake or wide river is the land along the edge of it.	0.588
57	A forest is a large area where trees grow close together. Woodland is land with a lot of trees.	0.628
58	An implement is a tool or other piece of equipment. A tool is any instrument or simple piece of equipment that you hold in your hands and use to do a particular kind of work.	0.590
59	A cock is an adult male chicken. A rooster is an adult male chicken.	0.863
60	A boy is a child who will grow up to be a man. A lad is a young man or boy.	0.580
61	A cushion is a fabric case filled with soft material, which you put on a seat to make it more comfortable.	0.523

	A pillow is a rectangular cushion which you rest your head on when you are in bed.	
62	A cemetery is a place where dead people's bodies or their ashes are buried. A graveyard is an area of land, sometimes near a church, where dead people are buried.	0.773
63	An automobile is a car. A car is a motor vehicle with room for a small number of passengers.	0.558
64	Midday is 12 o'clock in the middle of the day. Noon is 12 o'clock in the middle of the day.	0.955
65	A gem is a jewel or stone that is used in jewellery. A jewel is a precious stone used to decorate valuable things that you wear, such as rings or necklaces.	0.653

Table 5: the benchmark used for the second experiment [18]

ID	**Sentence**	**human assessed similarity**
66	Would you like to go out to drink with me tonight? I really dont know what to eat tonight so I might go out somewhere.	0.252
69	Does music help you to relax, or does it distract you too much? Does this sponge look wet or dry to you?	0.025
74	This key doesnt seem to be working, could you give me another? I dislike the word quay, it confuses me, I always think of things for locks, theres another one.	0.180
75	The ghost appeared from nowhere and frightened the old man. The ghost appeared from nowhere and frightened the old man.	0.362
77	The children crossed the road very safely thanks to the help of the lollipop lady. It was feared that the child might not recover, because he was seriously ill.	0.032
84	It seems like Ive got eczema on my ear doctor, can you recommend something for me? I had to go to a chemist for a special rash cream for my ear.	0.512

85	I am proud of our nation, well, most of it. I think of myself as being part of a nation.	0.427
86	There was a heap of rubble left by the builders outside my house this morning. Sometimes in a large crowd accidents may happen, which can cause deadly injuries.	0.020
92	Because I am the eldest one I should be more responsible. Just because of my age, people shouldnt think Im a responsible adult, but they do?	0.580
93	I need to dash into the kitchen because I think my chip pan is on fire. In the event of a chip pan fire follow the instructions on the safety note.	0.425
94	Peter was a very large youth, whose size intimidated most people, much to his delight. Now I wouldnt say he was fat, but Id certainly say he was one of the larger boys.	0.490
98	He was harshly punished for setting the fire alarms off. He delayed his response, in order to create a tense atmosphere.	0.055
99	Midday is 12 oclock in the middle of the day. Noon is 12 oclock in the middle of the day.	0.99
100	Thats not a very good car, on the other hand mine is great. This is a terrible noise level for a new car.	0.262
102	After hours of getting lost we eventually arrived at the hotel. After walking against the strong wind for hours he finally returned home safely.	0.272
105	Im worried most seriously about the presentation, not the essay. It is mostly very difficult to gain full marks in todays exam.	0.192
108	If you dont console with a friend, there is a chance you may hurt their feelings. One of the qualities of a good friend is the ability to console.	0.752
111	They said they were hoping to go to America on holiday. I like to cover myself up in lots of layers, I dont like the cold.	0.040
112	Will I have to drive far to get to the nearest petrol station?	0.960

	Is it much farther for me to drive to the next gas station?	
114	I am sorry but I cant go out as I have a heap of work to do. Ive a heap of things to finish so I cant go out Im afraid.	0.900
115	The responsible man felt very guilty when he crashed into the back of someones car. A slow driver can be annoying even though they are driving safely.	0.220
117	He fought in the war in Iraq before being killed in a car crash. The prejudice I suffered whilst on holiday in Iraq was quite alarming.	0.137
118	The cat was hungry so he went into the back garden to find lunch. The hen walked about in the yard eating tasty grain.	0.300
119	My bedroom wall is lemon coloured but my mother says it is yellow. Roses can be different colours, it has to be said red is the best though.	0.170
122	Flies can also carry a lot of disease and cause maggots. I dry my hair after I wash it or I will get ill.	0.030
125	The perpetrators of war crimes are rotten to the core. There are many global issues that everybody should be aware of, such as the threat of terrorism.	0.237
128	I hope youre taking this seriously, if not you can get out of here. The difficult course meant that only the strong would survive.	0.125
130	I bought a new guitar today, do you like it? The weapon choice reflects the personality of the carrier.	0.040
131	I am so hungry I could eat a whole horse plus dessert. I could have eaten another meal, Im still starving.	0.765

References

1. Marco De Boni and Suresh Manandhar. The use of sentence similarity as a semantic relevance metric for question answering. In *New Directions in Question Answering,*

Papers from 2003 AAAI Spring Symposium, Stanford University, Stanford, CA, USA, pages 138–144, 2003.
2. Harold J. Breaux. A modification of efroymson's technique for stepwise regression analysis. *Commun. ACM*, 11(8):556–558, August 1968.
3. Alexander Budanitsky and Graeme Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Comput. Linguist.*, 32(1):13–47, March 2006.
4. D. Croft, S. Coupland, J. Shell, and S. Brown. A fast and efficient semantic short text similarity metric. In *Computational Intelligence (UKCI), 2013 13th UK Workshop on*, pages 221–227, Sept 2013.
5. Marie-Catherine de Marneffe and Christopher D. Manning. The stanford typed dependencies representation. In *Coling 2008: Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, CrossParser '08, pages 1–8, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.
6. James W. Hardin and Joseph Hilbe. *Generalized Linear Models and Extensions*. College Station, Texas: Stata Press, 2001.
7. Vasileios Hatzlvassiloglou, Judith L. Klavans, and Eleazar Eskin. Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 203–212, 1999.
8. Virginia Heidinger. *Analyzing Syntax and Semantics: Workbook*. Gallaudet University Press, 1984.
9. G. Hirst and D. St Onge. *Lexical Chains as representation of context for the detection and correction malapropisms*. The MIT Press, May 1998.
10. Graeme Hirst and David St-Onge. WORDNET: A lexical database for English. In *Human Language Technology, Proceedings of a Workshop held at Plainsboro, New Jerey, USA, March 8-11, 1994*, 1994.
11. Aminul Islam and Diana Inkpen. Semantic text similarity using corpus-based word similarity and string similarity. *ACM Trans. Knowl. Discov. Data*, 2(2):10:1–10:25, July 2008.
12. Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition, 2000.
13. Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. An Introduction to Latent Semantic Analysis. *Discourse Processes*, (25):259–284, 1998.
14. Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. Dbpedia - A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195, 2015.
15. Yuhua Li, David McLean, Zuhair A. Bandar, James D. O'Shea, and Keeley Crockett. Sentence similarity based on semantic nets and corpus statistics. *IEEE Trans. on Knowl. and Data Eng.*, 18(8):1138–1150, August 2006.
16. Lee Ming Che, Chang Jia Wei, Hsieh Tung Cheng, Chen Hui Hui, and Ching Hui Chen. A sentence similarity metric based on semantic patterns. *Advances in Information Sciences and Service Sciences.*, 4(1):576–585, October 2012.
17. Jesus Oliva, Jose Ignacio Serrano, M. Dolores del Castillo, and Ángel Iglesias. Symss: A syntax-based measure for short-text semantic similarity. *Data Knowl. Eng.*, 70(4):390–405, 2011.
18. James O'shea, Zuhair Bandar, and Keeley Crockett. A new benchmark dataset with production methodology for short text semantic similarity algorithms. *ACM Trans. Speech Lang. Process.*, 10(4):19:1–19:63, January 2014.
19. James O'Shea, Zuhair Bandar, Keeley A. Crockett, and David McLean. A comparative study of two short text semantic similarity measures. In *Agent and Multi-Agent Systems: Technologies and Applications, Second KES International Symposium, KES-AMSTA 2008, Incheon, Korea, March 26-28, 2008. Proceedings*, pages 172–181, 2008.
20. Rakesh Peter, G Shivapratap, G Divya, and KP Soman. Evaluation of svd and nmf methods for latent semantic analysis. *International Journal of Recent Trends in Engineering*, 1(3), 2009.

21. Gerard Salton and Michael McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company, 1984.
22. Gerard Salton, Andrew Wong, and Chungshu Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, November 1975.
23. Alexandre Spaeth and Michel C. Desmarais. Combining collaborative filtering and text similarity for expert profile recommendations in social websites. In *User Modeling, Adaptation, and Personalization - 21th International Conference, UMAP 2013, Rome, Italy, June 10-14, 2013, Proceedings*, pages 178–189, 2013.
24. George Tsatsaronis, Iraklis Varlamis, and Michalis Vazirgiannis. Text relatedness based on a word thesaurus. *J. Artif. Int. Res.*, 37(1):1–40, January 2010.
25. William E. Winkler. The state of record linkage and current research problems. Technical report, Statistical Research Division, U.S. Census Bureau, 1999.