Pilot Short Text Semantic Similarity Benchmark Data Set: Full Listing and Description

James O'Shea*, Zuhair Bandar, Keeley Crockett, David McLean

Department of Computing and Mathematics, Manchester Metropolitan University, Chester St., Manchester M1 5GD, United Kingdom
(j.d.oshea, z.Bandar, k.crockett, d.mclean)@mmu.ac.uk

*Corresponding Author for information about this data set

Abstract

This report contains a listing of all of the sentence pairs comprising the pilot benchmark data set for evaluating algorithms designed to measure Short Text Semantic Similarity (STSS). A short text is a coherent piece of text at the sentence level, but which does not necessarily conform to the grammatical rules of correctly formed sentences. Thus it includes spoken and typed utterances. This benchmark data set has been used in publications concerning our own STSS measure, now known as STASIS and has been requested by other scientists working in the field. Because length restrictions have prevented it from being included in full in our other publications, we make it available through this technical report.

Keywords

Conversational Agents, Sentence similarity, term similarity, text similarity, semantic similarity, semantic distance, semantic relatedness, natural language processing

Issue date: 2009-07-02

Citation information

Please cite [1], the paper in which this data set was first used. For a fuller understanding of how the data set was collected and the method for using it to compare STSS measures please see [2] which you may also wish to cite.
Acknowledgement: The sentences used in this data set are taken from the Collins Cobuild Dictionary[3] with some minor modifications described later.

Rated Similarity of Sentence Pairs

Mean & Standard Deviation of ratings provided by 32 human participants

Sentence pair	$\overline{X}$	s
1.cord:smile 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.	0.04	0.16
2.rooster:voyage A rooster is an adult male chicken. A voyage is a long journey on a ship or in a spacecraft.	0.02	0.09
3.noon:string Noon is 12 o'clock in the middle of the day. String is thin rope made of twisted threads, used for tying things together or tying up parcels.	0.05	0.2
4.fruit:furnace 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. 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.	0.19	0.63
5.autograph:shore 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.	0.02	0.09
6.automobile:wizard An automobile is a car. In legends and fairy stories, a wizard is a man who has magic powers.	0.08	0.26
7.mound:stove A mound of something is a large rounded pile of it. A stove is a piece of equipment which provides heat, either for cooking or for heating a room.	0.02	0.26
8.grin:implement A grin is a broad smile. An implement is a tool or other piece of equipment.	0.02	0.08
9.asylum:fruit 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.	0.02	0.09
10.asylum:monk An Asylum is a psychiatric hospital. A monk is a member of a male religious community that is usually separated from the outside world.	0.15	0.34

Sentence pair	$\overline{X}$	s
11.graveyard:madhouse A graveyard is an area of land, sometimes near a church, where dead people are buried. If you describe a place or situation as a madhouse you mean that it is full of confusion and noise.	0.09	0.32
12.glass:magician Glass is a hard transparent substance that is used to make things such as windows and bottles. A magician is a person who entertains people by doing magic tricks.	0.03	0.12
13.boy:rooster A boy is a child who will grow up to be a man. A rooster is an adult male chicken.	0.43	0.72
14.cushion:jewel A cushion is a fabric case filled with soft material, which you put on a seat to make it more comfortable. A jewel is a precious stone used to decorate valuable things that you wear, such as rings or necklaces.	0.21	0.48
15.monk:slave A monk is a member of a male religious community that is usually separated from the outside world. A slave is someone who is the property of another person and has to work for that person.	0.18	0.38
16.asylum:cemetery An Asylum is a psychiatric hospital. A cemetery is a place where dead people's bodies or their ashes are buried.	0.15	0.42
17.coast:forest The coast is an area of land that is next to the sea. A forest is a large area where trees grow close together.	0.19	0.34
18.grin:lad A grin is a broad smile. A lad is a young man or boy.	0.05	0.14
19.shore:woodland The shores or shore of a sea, lake or wide river is the land along the edge of it. Woodland is land with a lot of trees.	0.33	0.42
20.monk:oracle A monk is a member of a male religious community that is usually separated from the outside world. In ancient times, an oracle was a priest or priestess who made statements about future events or about the truth.	0.45	0.61
21.boy:sage 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.	0.17	0.35

Sentence pair	$\overline{X}$	s
22.automobile:cushion An automobile is a car. A cushion is a fabric case filled with soft material, which you put on a seat to make it more comfortable.	0.08	0.21
23.mound:shore A mound of something is a large rounded pile of it. The shores or shore of a sea, lake or wide river is the land along the edge of it.	0.14	0.29
24.lad:wizard A lad is a young man or boy. In legends and fairy stories, a wizard is a man who has magic powers.	0.13	0.27
25.forest:graveyard 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.	0.26	0.4
26.food:rooster Food is what people and animals eat. A rooster is an adult male chicken.	0.22	0.44
27.cemetery:woodland A cemetery is a place where dead people's bodies or their ashes are buried. Woodland is land with a lot of trees.	0.15	0.31
28.shore:voyage The shores or shore of a sea, lake or wide river is the land along the edge of it. A voyage is a long journey on a ship or in a spacecraft.	0.08	0.23
29.bird:woodland 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.	0.05	0.18
30.coast:hill The coast is an area of land that is next to the sea. A hill is an area of land that is higher than the land that surrounds it.	0.40	0.53
31.furnace:implement 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. An implement is a tool or other piece of equipment.	0.20	0.46
32.crane:rooster A crane is a large machine that moves heavy things by lifting them in the air. A rooster is an adult male chicken.	0.08	0.22

Sentence pair	$\overline{X}$	s
33.hill:woodland A hill is an area of land that is higher than the land that surrounds it. Woodland is land with a lot of trees.	0.58	0.73
34.car:journey A car is a motor vehicle with room for a small number of passengers. When you make a journey, you travel from one place to another.	0.29	0.52
35.cemetery:mound A cemetery is a place where dead people's bodies or their ashes are buried. A mound of something is a large rounded pile of it.	0.23	0.44
36.glass:jewel Glass is a hard transparent substance that is used to make things such as windows and bottles. A jewel is a precious stone used to decorate valuable things that you wear, such as rings or necklaces.	0.43	0.71
37.magician:oracle 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.	0.52	0.65
38.crane:implement A crane is a large machine that moves heavy things by lifting them in the air. An implement is a tool or other piece of equipment.	0.74	0.89
39.brother:lad Your brother is a boy or a man who has the same parents as you. A lad is a young man or boy.	0.51	0.53
40.sage:wizard A sage is a person who is regarded as being very wise. In legends and fairy stories, a wizard is a man who has magic powers.	0.61	0.78
41.oracle:sage 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.	1.13	0.94
42.bird:crane A bird is a creature with feathers and wings, females lay eggs and most birds can fly. A crane is a large machine that moves heavy things by lifting them in the air.	0.14	0.3

Sentence pair	$\overline{X}$	s
43.bird:cock A bird is a creature with feathers and wings, females lay eggs and most birds can fly. A cock is an adult male chicken.	0.65	0.63
44.food:fruit Food is what people and animals eat. 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.	0.97	0.94
45.brother:monk Your brother is a boy or a man who has the same parents as you. A monk is a member of a male religious community that is usually separated from the outside world.	0.18	0.37
46.asylum:madhouse An Asylum is a psychiatric hospital. If you describe a place or situation as a madhouse you mean that it is full of confusion and noise.	0.86	0.86
47.furnace:stove 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.	1.39	0.99
48.magician:wizard 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	1.42	0.92
49.hill:mound 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.	1.17	1.01
50.cord:string Cord is strong, thick string. String is thin rope made of twisted threads, used for tying things together or tying up parcels.	1.88	1.22
51.glass:tumbler 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.	0.55	0.68
52.grin:smile 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.	1.94	1.29

Sentence pair	$\overline{X}$	s
53.serf:slave 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.	1.93	1.09
54.journey:voyage 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.	1.44	1.09
55.autograph:signature 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.	1.62	1.29
56.coast:shore 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.	2.35	1.05
57.forest:woodland A forest is a large area where trees grow close together. Woodland is land with a lot of trees.	2.51	1.05
58.implement:tool 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.	2.36	1.28
59.cock:rooster A cock is an adult male chicken. A rooster is an adult male chicken.	3.45	0.81
60.boy:lad A boy is a child who will grow up to be a man. A lad is a young man or boy.	2.32	1.24
61.cushion:pillow A cushion is a fabric case filled with soft material, which you put on a seat to make it more comfortable. A pillow is a rectangular cushion which you rest your head on when you are in bed.	2.09	1.06
62.cemetery: graveyard 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.	3.09	0.92

Sentence pair	$\overline{X}$	s
63.automobile:car An automobile is a car. A car is a motor vehicle with room for a small number of passengers.	2.23	1.26
64.midday:noon Midday is 12 o'clock in the middle of the day. Noon is 12 o'clock in the middle of the day.	3.82	0.54
65.gem: jewel 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.	2.61	1.31

Jim O'Shea's notes on using the data set.

I created this set as part of my PhD (supervised by Bandar & Crockett) as a benchmark for validating STSS measures. The immediate need arose from our research group, to evaluate the STASIS measure (and improvements made to it in the future). It is also intended to be used by the research community at large, to evaluate and compare measures arising from future work in this area. A further set of 66 sentence pairs is under development which is intended to cover a greater semantic space and represent a wider range of speech acts. It will appear in the literature and on this website in due course.
You are invited to use this data set, citing the source papers mentioned above. The following guidance is intended to help make benchmark tests performed with the data set comparable.

1. An STSS measure can be validated by comparing its performance with human ratings, in particular the ratings that a "typical" human might give.

2. The ratings in the table follow the practice used in word similarity studies[4]. The "typical" human rating is the mean of those given by a set of participants. The measure of agreement is the Pearson product-moment correlation coefficient (r) quoted with statistical significance.

3. To benefit from the experience of earlier work, this pilot data set uses the 65 word pairs created by Rubenstein & Goodenough [5], replacing the individual words with their definitional sentence from the Collins Cobuild Dictionary [3]. Some of the definitions were modified and this is explained later.

4. A property that the data set inherits from the original word pairs is a heavy bias towards low similarity pairs and this could bias the validation of STSS measures. Consistent with approaches to word similarity measurement [1, 4], we selected a subset with a more representative distribution across the range. This subset is discussed in [2] and is distinguished here by showing the sentence number in **bold** type (e.g. **64.**midday:noon). It is essential to use this same subset to make a meaningful comparison of a new STSS with those already tested using the dataset.

5. The ratings are the numbers in the $\overline{X}$ column; they are from a rating scale running from 0.00 to 4.00. The simplest procedure is to calculate the correlation coefficient between a new measure and the human ratings in the original range (0.00 – 4.00). Linear transformations are permissible, e.g. dividing by 4 to re-scale them to run from 0.00 to 1.00 to compare the ratings for STASIS with LSA. Re-scaling should not lead to a different correlation co-efficient (however, see below on rounding noise).

6. Degree of accuracy. Ratings are collected to 2 significant digits and the means are calculated to 3 significant digits, in keeping with the "estimated digit" used in physical science measurements. As the absolute maximum value of *r* is 1.000, (which would indicate perfect agreement between the algorithm and the "typical" human), we take the maximum accuracy for quoting *r* as 3 decimal places.

7. Consistency with earlier studies. Applying different rounding procedures can introduce noise and lead to variations in the least significant digit of *r*. For comparability with [1] and [2], perform the linear transformation on the human ratings then round them. Calculate *r*, then round *r* to 3 decimal places.
Common sense dictates that as the least significant digit of the 3 is based on the estimated digit, the importance of differences between measures based on this digit alone should not be exaggerated.

8. Those familiar with measurement theory may argue that mean and *r* are unsuitable statistics for data collected on this measurement scale. We are aware of the argument; however we have used the techniques because they are well-established and understood in word similarity. Furthermore in [2] we describe the data collection process and the steps taken to improve ratio scale properties.

9. Modifications to the Cobuild definitions.

Two of the original Cobuild sentences were modified. These were smile and bird.

The definition of smile is circular:

A smile is the expression that you have on your face when you smile.

The alternative substitutes in a fragment from the definition of the verb smile:

A smile is the expression that you have on your face when you are pleased or amused, or when you are being friendly.

The definition of bird spans 3 sentences, all of which contribute.

A bird is a creature with feathers and wings. (Female birds lay eggs. Most birds can fly.)

The alternative incorporates these as clauses in a single sentence:

A bird is a creature with feathers and wings, female birds lay eggs and most can fly.

Comments on other definitions:

Rooster/cock and Midday/Noon are genuinely identical definitions and have been left unmodified.

Automobile
The definition of automobile:

An automobile is a car.

Has not been modified (the obvious change would be to substitute in the definition of car). However, car is a member of the data set and is paired with automobile. It is an interesting combination to investigate – should humans rate them as identical? It actually came out as medium similarity, with a high standard deviation indicating quite a lot of uncertainty on the part of the human raters.

10. A minor typographical error resulted in an incorrect similarity value for Sentence Pair 17 in [1] and [2]. This is corrected in table above. When the correlation coefficients are re-calculated the value of *r* for STASIS increases from 0.816 to 0.817, the value for LSA remains unchanged.

Further reading
Other useful material may be found in Rubenstein & Goodenough[5], Oppenheim [6] and Li et Al [7].

References

1. Li, Y., et al., *Sentence Similarity Based on Semantic Nets and Corpus Statistics.* IEEE Transactions on Knowledge and Data Engineering, 2006. **18**(8): p. 1138-1150.
2. O'Shea, J., et al., *A Comparative Study of Two Short Text Semantic Similarity Measures* Lecture Notes in Artificial Intelligence, 2008. **4953**.
3. Sinclair, J., *Collins Cobuild English Dictionary for Advanced Learners*. 3 ed. 2001: HarperCollins.
4. Miller, G.A. and W.G. Charles, *Contextual Correlates of Semantic Similarity.* Language and Cognitive Processes 1991. **6** (1): p. 1-28.
5. Rubenstein, H. and J. Goodenough, *Contextual Correlates of Synonymy,* Communications of the ACM, 1965. **8**(10): p. 627-633.
6. Oppenheim, A.N., *Questionnaire Design, Interviewing and Attitude Measurement*. 1992: Continuum
7. Li, Y., Z. Bandar, and D. McLean, *An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources.* IEEE Transactions on Knowledge and Data Engineering, 2003, **15**(4).