

---

# Exploring Semantic Textual Similarity

---

*Master thesis by*

**Aitor Gonzalez Agirre**

Advisors

Dr. Eneko Agirre Bengoa

Dr. German Rigau Claramunt

**Master in Analysis and Processing of Language**

Dept. of Computer Languages and Systems

Dept. of Computer Architecture and Technology

University of the Basque Country (UPV/EHU)

Donostia-San Sebastián, September 2012



# Abstract

Measuring semantic similarity and relatedness between textual items (words, sentences, paragraphs or even documents) is a very important research area in Natural Language Processing (NLP). In fact, it has many practical applications in other NLP tasks. For instance, Word Sense Disambiguation, Textual Entailment, Paraphrase detection, Machine Translation, Summarization and other related tasks such as Information Retrieval or Question Answering.

In this master thesis we study different approaches to compute the semantic similarity between textual items. In the framework of the european PATHS project<sup>1</sup>, we also evaluate a knowledge-base method on a dataset of cultural item descriptions. Additionally, we describe the work carried out for the Semantic Textual Similarity (STS) shared task of SemEval-2012. This work has involved supporting the creation of datasets for similarity tasks, as well as the organization of the task itself.

---

<sup>1</sup><http://www.paths-project.eu/>



# Acknowledgements

Firstly, I would like to thank the IXA NLP group from the Basque Country University. This work was been possible thanks to its support withing the framework of the KNOW2 (TIN2009-14715-C04-04) and PATHS (FP7-ICT-2009-6-270082) projects.

Finally, I would like to thank Eneko Agirre and German Rigau. Whenever I needed, they provided their help, and my hope is to continue that way.



# List of Figures

3.1	Example of a Europeana Semantic Element . . . . .	14
3.2	Europeana item . . . . .	15
3.3	Interface used to obtain human judgments . . . . .	16
3.4	Culture Grid: Averages of the results with different knowledge bases. . . . .	24
3.5	SCRAN: Averages of the results with different knowledge bases.	25
3.6	Culture Grid: Averages of the results with UKB's similarity modes (cos and dot). . . . .	26
3.7	SCRAN: Averages of the results with UKB's similarity modes (cos and dot). . . . .	27
3.8	Culture Grid: Averages of the results obtained with the different features combinations. . . . .	28
3.9	SCRAN: Averages of the results obtained with the different features combinations. . . . .	28
3.10	Culture Grid: Maximum, mean and minimum obtained by Spearman and Pearson correlations. . . . .	29
3.11	SCRAN: Maximum, mean and minimum obtained by Spearman and Pearson correlations. . . . .	29
4.1	Video and corresponding descriptions from MSRvid . . . . .	33
4.2	Definition and instructions for annotation . . . . .	34





# Contents

Abstract . . . . .	i
Acknowledgments . . . . .	iii
List of Figures . . . . .	v
Contents . . . . .	vii
<b>1 Introduction</b>	<b>1</b>
1.1 Document structure . . . . .	2
<b>2 State of the Art</b>	<b>3</b>
2.1 Semantic Similarity . . . . .	3
2.2 Measuring Semantic Similarity . . . . .	4
2.3 Knowledge-based similarity . . . . .	4
2.3.1 WordNet-based methods . . . . .	5
2.3.2 Wikipedia-based methods . . . . .	7
2.4 Corpus-based similarity . . . . .	8
2.4.1 Distributional Semantics . . . . .	8
2.5 Combining Knowledge-based and Corpus-based similarity . .	9
2.6 Datasets . . . . .	9
2.6.1 RG dataset . . . . .	10
2.6.2 WordSim-353 dataset . . . . .	10
2.6.3 Li and Lee datasets . . . . .	11
2.7 Conclusions . . . . .	12
<b>3 Applying similarity within PATHS project</b>	<b>13</b>
3.1 Introduction . . . . .	13
3.2 Europeana . . . . .	13
3.3 Creating a dataset . . . . .	14
3.3.1 Human judgments on the dataset . . . . .	16
3.4 Computing similarity using random walks . . . . .	17
3.4.1 Experiment setup . . . . .	18
3.5 Results . . . . .	18
3.5.1 Using WordNet 3.0 . . . . .	20
3.5.2 Using WordNet 3.0 enriched with gloss relations . . .	20
3.5.3 Using WordNet 3.0 enriched with gloss relations and KnowNet-5 . . . . .	22

---

3.5.4	Using pre-calculated vectors from WordNet 3.0 enriched with gloss relations . . . . .	22
3.5.5	Comparison of the obtained results . . . . .	23
3.5.6	Comparison with the state of the art . . . . .	25
3.6	Conclusions . . . . .	26
<b>4</b>	<b>SemEval-2012 Task 6</b>	<b>31</b>
4.1	Introduction . . . . .	31
4.2	Source Datasets . . . . .	32
4.3	Annotation . . . . .	35
4.4	Systems Evaluation . . . . .	36
4.4.1	Evaluation metrics . . . . .	36
4.4.2	Using confidence scores . . . . .	37
4.4.3	The Baseline System . . . . .	37
4.4.4	Participation . . . . .	37
4.5	Results . . . . .	38
4.6	Tools and Resources used . . . . .	38
4.7	Best three systems . . . . .	39
4.7.1	baer/run2 . . . . .	39
4.7.2	jan snajder/run1 . . . . .	39
4.7.3	sgjimenezv/run1 . . . . .	40
4.8	Our system . . . . .	40
4.9	Conclusions . . . . .	41
<b>5</b>	<b>Concluding Remarks and Future Directions</b>	<b>45</b>
5.1	Main contributions . . . . .	45
5.2	Future work . . . . .	46
	<b>Bibliography</b>	<b>47</b>

# Chapter 1

## Introduction

Measuring semantic similarity and relatedness between textual items (words, sentences, paragraphs or even documents) is a very important research area in Natural Language Processing (NLP). In fact, it has many practical applications in other NLP tasks. For instance, Word Sense Disambiguation [Li et al., 1995], Textual Entailment [Castillo and Cardenas, 2010], Paraphrase detection [Fern and Stevenson, 2009], Machine Translation [Banchs and Costa-jussà, 2011], Summarization [Nagwani and Verma, 2011] and other related tasks such as Information Retrieval [Hliaoutakis et al., 2006] or Question Answering [Mohler et al., 2011].

The techniques used to solve this problem can be roughly classified into two main categories. On the one hand, those relying on pre-existing knowledge resources (thesauri, semantic networks, taxonomies or encyclopedias) [Alvarez and Lim, 2007, Yang and Powers, 2005, Hughes and Ramage, 2007, Agirre and Soroa, 2009a]. On the other hand, those inducing distributional properties from corpora [Sahami and Heilman, 2006, Chen et al., 2006, Bollegala et al., 2009, Agirre and Soroa, 2009a].

However, despite the large amount of available techniques this is still an open and very active research area. The following examples can help us to illustrate the difficulty of the task.<sup>1</sup>

- *Fred saw the train flying over Berna.*
- *Fred saw the plane flying over Berna.*

The above both sentences look quite similar. Just one word changes totally their meaning. They represent two different realities. In the first sentence, it seems that Fred is flying over Berna on an airplane (or another flying vehicle) and from there he is watching a train. In the second sentence, it seems that Fred is out of the plane which is flying over Berna.

- *Fred saw the plane flying over Berna.*
- *Fred watched the jet soaring over the capital of Switzerland.*

---

<sup>1</sup>The first one inspired from an example of Cyc <http://www.cyc.com/>

Instead, the previous two sentences are very similar even though they are realized very differently. It only seems that the second sentence is more precise than the first one. But both meanings are compatible.

- *Fred saw the plane flying over Berna.*
- *Fred vió el avión volando sobre Berna.*

In this case, although both sentences are in different languages (English and Spanish), both sentences provide the same meaning.

## 1.1 Document structure

After this short introduction, Chapter 2 provides an in depth review of different methods to compute the semantic similarity between textual items.

Chapter 3 presents our work in the framework of the european PATHS project<sup>2</sup>. In this framework, we evaluate a knowledge-based semantic similarity method on a dataset of cultural item descriptions obtained from Europeana<sup>3</sup>. Europeana is a internet portal that acts as an interface for large collections of cultural items. It contains millions of books, paintings, films and other museum objects that have been digitized throughout Europe. It is an on-line resource that stores the cultural heritage of Europe. Users are able to explore these collections which contain images, sounds and video from different providers. Although it provides advanced searching and browsing facilities, there is no sufficient exploitation of the semantic content of these multilingual collections. For example, each item includes text features such as title, description and subject. Computing the textual semantic similarity between them would be useful to help users navigate through this vast resource. In this Master thesis we apply a knowledge-based method to compute the similarity between Europeana items. We evaluate these results using a gold-standard created by human judgments.

Chapter 4 describes the work carried out for the Semantic Textual Similarity (STS) shared task of SemEval-2012. This work has involved supporting the creation of datasets for similarity tasks, as well as the organization of the task itself. This work also resulted in a publication [Agirre et al., 2012].

Finally, in Chapter 5 we drawn some concluding remarks and provide some lines for future research.

---

<sup>2</sup><http://www.paths-project.eu/>

<sup>3</sup><http://www.europeana.eu>

## Chapter 2

# State of the Art

This chapter provides a revision of the state-of-the-art on computational lexical semantics for Natural Language Processing (NLP) and presents several methods proposed for computing semantic similarity between words or texts.

### 2.1 Semantic Similarity

Lexical semantics is the area of NLP that studies the meaning of the words. Of course, there are many words that have more than one meaning. Consider as an example the following sentences:

- *I traveled by train from Barcelona to Donostia.*
- *Mikel and Joseba train every day at the gym.*

The word **train** has different meaning in the above sentences:

- *a series of connected railway carriages or wagons moved by a locomotive or by integral motors.*
- *to make (a person) fit by proper exercise, diet, practice, etc., as for an athletic performance.*

Therefore, we can say that there are words that are spelled in the same way but with different senses. Basically, a sense is one of the possible meanings of a given word. If two different senses of a word are not semantically related between them we are talking about a homonymy relation, as the example with train we just saw. Instead, if two senses of a word are semantically related we are talking about polysemy. Consider the example of the word **wood**:

- *a piece of a tree.*
- *a geographical area with many trees.*

Finally, we can say that concepts that share some meaning are semantically similar. For example, *dog* and *cat* are more semantically related than *house* and *train*. But if we compare *dog* and *cat* with *car* and *bus* the thing is not so clear: *cat* and *dog* are pets, and both *car* and *bus* are on wheels means of transportation, so that both pairs of words are very related between them.

## 2.2 Measuring Semantic Similarity

It is commonly accepted that there are at least two kinds of methods to determine whether two words, phrases or texts share some kind of meaning. The first one is based on structured resources such as monolingual or bilingual dictionaries, thesaurus or encyclopedias. These structured resources are very useful because they constitute a highly structured and relevant source of information about words and meanings. Some of the more employed resources of this type are WordNet [Fellbaum, 1998a] and Wikipedia<sup>1</sup>. This kind of algorithms often use the hypernym/hyponym relations (e.g. in WordNet) to compute the semantic between two concepts. These types of resources are more detailed in Section 2.3.

Large corpora has been also used as a source data for semantic similarity. The possibility of applying *descriptive approaches* using statistical techniques, having information of the frequency of use, etc. is crucial for extracting important information related to linguistic phenomena. Thus, *unstructured lexical resources* such as monolingual and bilingual corpora provide an additional though less organized source for semantic similarity. The most widely used representation of the features in a document (or corpus) is the Vector Space model [Salton et al., 1975].

Most of these techniques are applied at word level, and very few at sentence level. This is because compositionality, which makes calculating the similarities between sentences very complex and difficult. In Chapter 4 we try to deal with this problem.

## 2.3 Knowledge-based similarity

In NLP, the use of *on-line dictionaries* or Machine Readable Dictionaries (MRDs), a term coined in the 80s referring to dictionaries for human use in digital support, has been studied extensively in the hope that monolingual and bilingual dictionaries might provide a way out of the semantic similarity. Although MRDs are built for human use and they deal with problems such as inconsistencies, too fine-grained ambiguity, circular definitions, etc., MRDs seemed to offer the possibility for enormous savings in time and human effort [Zernik, 1991, Briscoe and Boguraev, 1989, Wilks et al., 1996, Rigau et al., 1998].

---

<sup>1</sup><http://www.wikipedia.org>

### 2.3.1 WordNet-based methods

One of the most important and popular knowledge-bases is WordNet. This section illustrates some of the best known techniques based on WordNet that allows us to calculate the similarity between concepts:

- **Path-Length Measure:** This algorithm is based on the principal assumption that the shorter the path between two words is, more similar they are between them.
- **Leacock-Chodorow Measure:** This method is an extension to the Path-Length measure which scales the path length by the depth of the hierarchy, defined as the length of the longest path from a leaf node to the root of the hierarchy [Leacock and Chodorow, 1998].
- **Resnik Similarity Measure:** This algorithm uses the structure of the thesaurus and combines it with probabilistic information extracted from corpora. Resnik's similarity measure supposes that the semantic similarity of two concepts is proportional to the amount of information they share [Resnik, 1995].
- **Lin Similarity Measure:** is an extension the Resnik similarity, introducing the *commonality* and *difference* measures. *Commonality* is a measure that indicates how much two concepts have in common. *Difference* is the measure that indicates that the more differences are between two concepts, the more different they are [Lin, 1997].
- **Jiang-Conrath Distance:** This technique measures unrelatedness between two concepts [Jiang and Conrath, 1997].
- **Hirst-St-Onge Measure:** The algorithm classifies the WordNet relations in three categories: up, down or horizontal. There are also four levels of relatedness: extra strong, strong, medium strong and weak. The extra strong and strong relationship involve words of the same concept (horizontal relation). [Hirst and St-Onge, 1998] calculates the score of the relation with the path length between the concepts and the number of changes of direction in that path.

Moreover, [Pedersen and Patwardhan, 2004] created a freely available software package that makes it possible to measure the semantic similarity and relatedness between a pair of concepts (or synsets). It provides six measures of similarity, and three measures of relatedness, all of which are based on WordNet. These measures are implemented as Perl modules called *WordNet::Similarity* which take as input two concepts, and return a numeric value that represents the degree to which they are similar.

#### 2.3.1.1 Extended Lesk Measure

The Lesk Algorithm [Lesk, 1986] is an algorithm based on two assumptions. The first one is that concepts that are nearby between them have more

possibilities to share some topic. The second is that related senses can be identified searching overlaps in their glosses.

The algorithm computes simple unigram overlaps in the glosses that are contained in WordNet. The basic idea behind the Extended Lesk measure [Patwardhan et al., 2003] is that two concepts in a dictionary are similar if they share common words in their glosses. For each common phrase in the glosses of two concepts containing  $n$  words, the Extended Lesk measure assigns a score of  $n^2$ . The total similarity score is the sum of those scores. In addition, Extended Lesk looks for overlap between all glosses of the senses that have a relation (e.g. hypernym, hyponym) with the concepts.

Let  $R$  be the set of possible WordNet relations between two concepts. The Extended Lesk overlap measure is defined as:

$$sim_{eLesk}(c1, c2) = \sum_{r, q \in R} overlap(gloss(r(c1)), gloss(q(c2))) \quad (2.1)$$

Where  $c1$ ,  $c2$  are two concepts,  $r$ ,  $q$  are two WordNet relations and  $gloss(r(c))$  is the concatenation of all the senses of  $c$  with relation  $r$ .

### 2.3.1.2 Graph-based Method

This method considers WordNet as a graph  $G = (V, E)$  in which each node represent a concept (synset) or a dictionary word. Each undirected edge represents a relation between synsets and each directed edge represents a link from a dictionary word to a synset. [Hughes and Ramage, 2007] presented a random walk algorithm over WordNet, with good results on a similarity dataset. [Agirre et al., 2009] improved these results and provided the best results among WordNet-based algorithms on the Wordsim353 dataset.

The method includes two steps. Firstly, it computes a variant of the original PageRank [Page et al., 1999] called personalised PageRank [Haveliwala, 2002] over WordNet for each word in order to produce a probability distribution over WordNet synsets. Then, it computes the similarity of those words by using the cosine between two vectors created from the probability distributions.

In the first step,  $G$  is considered as a graph with  $N$  vertices  $v_1, \dots, v_N$  and  $d_i$  be the out-degree of node  $i$ ; let  $M$  be a  $N \times N$  transition probability matrix, where  $M_{ji} = \frac{1}{d_i}$  if a link from  $i$  to  $j$  exists, and zero otherwise. Then, the calculation of the *PageRank vector*  $\mathbf{Pr}$  over  $G$  is equivalent to resolving the following equation:

$$\mathbf{Pr} = cM\mathbf{Pr} + (1 - c)\mathbf{v} \quad (2.2)$$

In the equation,  $\mathbf{v}$  is a  $N \times N$  vector whose elements are  $\frac{1}{N}$  and  $c$  is the so called *damping factor*, a scalar value between 0 and 1. The first term of the sum on the equation models the voting scheme described in the beginning of the section. The second term represents, loosely speaking, the probability of



a surfer randomly jumping to any node, e.g. without following any paths on the graph. The damping factor, usually set in the  $[0.85..0.95]$  range, models the way in which these two terms are combined at each step.

In the second step, once personalized PageRank is computed, it returns a probability distribution over WordNet synsets. The similarity between two words can thus be implemented as the similarity between the probability distributions. Alternatively, we can interpret the probability distribution for a word  $w$  as a vector  $\vec{w}$  of weights  $w_i$  where each dimension  $i$  is a synset, and use the cosine to compute similarity, as in the following equation:

$$\text{similarity}(\vec{w}, \vec{v}) = \cos(\vec{w}, \vec{v}) = \frac{\vec{w} \cdot \vec{v}}{\|\vec{w}\| \|\vec{v}\|} \quad (2.3)$$

This method is implemented in the UKB<sup>2</sup> package, a collection of programs for performing graph-based Word Sense Disambiguation and lexical similarity/relatedness using a pre-existing knowledge base [Agirre et al., 2009, Agirre et al., 2010]. UKB has been developed by the IXA<sup>3</sup> group in the University of the Basque Country.

### 2.3.2 Wikipedia-based methods

Lately, a new approach has entered into the scene: building wide coverage knowledge bases from *encyclopedias* developed by Web2.0 communities, such as Wikipedia<sup>4</sup>. Wikipedia is a multilingual, Web-based encyclopedia written collaboratively by volunteers which is available for free. This section describes some methods based on Wikipedia that allows us to calculate the similarity between concepts:

- **WikiRelate!**: This system developed by [Strube and Ponzetto, 2006] is based on methods for WordNet [Hirst and St-Onge, 1998, Jiang and Conrath, 1997, Leacock and Chodorow, 1998, Lin, 1997, Patwardhan et al., 2003, Resnik, 1995] and redesigned to work with the Wikipedia. WikiRelate! retrieves all pages from Wikipedia containing the two words for which we want to compute the similarity, and then computes the text overlaps in the content of the articles.
- **Wikipedia Link Vector Model**: This technique is based in the structure of the links and the titles of the Wikipedia articles. The system computes the similarity computing the angle between the vectors of links, weighting them with the probability of each link.
- **WikiWalk**: WikiWalk [Yeh et al., 2009] is a method that uses random walk algorithms on a graph to measure semantic similarity between words. The graph is created by representing each article as a node and each link between articles as an edge. Given two words, WikiWalk uses

<sup>2</sup><http://ixa2.si.ehu.es/ukb>

<sup>3</sup><http://ixa.si.ehu.es/Ixa>

<sup>4</sup><http://www.wikipedia.org>

the Explicit Semantic Analysis [Gabrilovich and Markovitch, 2007] to find their corresponding nodes in the Wikipedia graph. After the words are linked to specific nodes, semantic similarity is computed by applying personalised Pagerank for each word to create a probability distribution of related nodes. The final score is given by the cosine of the angle between the vectors of their probability distributions.

## 2.4 Corpus-based similarity

Large corpora has been also used as a source data for semantic similarity. The possibility of applying *descriptive approaches* (those which derive the necessary knowledge from a natural source of data without any preexisting frame) using statistical techniques, having information of the frequency of use, etc. is crucial for extracting important information related to linguistic phenomena. Thus, *unstructured lexical resources* such as monolingual and bilingual corpora provide an additional though less organized source for semantic similarity.

### 2.4.1 Distributional Semantics

*Distributional Semantics Modelling* (DSM) is an active area of research within the field of natural language processing. In distributional semantics, the meaning of words is explored by looking at their use (their distribution) in texts. The combined contexts of words, represented as feature vectors in a high-dimensional vector space, are indicative of their meanings. This models are named Vector Space Models (VSM).

In VSM the meaning of a content word is represented in terms of a distributed vector, recording its pattern of co-occurrences (sometimes, using specific syntactic relations) with respect other content words within a corpus. Different semantic similarity measures and linguistic phenomena may then be modeled in terms of linear algebra operations (such as cosine) on distributional vectors.

Since distributional semantic models represent words according to their occurrence contexts, they may be used to model word similarity or word association (i.e. two words are similar/related if they co-occur in similar contexts). This idea can be straightforwardly used to acquire pairs (or sets) of related words.

Many studies have used different statistic techniques to measure the significance of terms with respect a corpus text. In Information Retrieval [Baeza-Yates and Ribeiro-Neto, 1999, Kageura and Umino, 1996, Manning and Schütze, 1998] different term-weight measures are used to represent the usefulness of terms in the retrieval process; for example, frequency [Luhn, 1957], signal-to-noise ratio [Dennis, 1964, Salton and McGill, 1986], IDF [Jones, 1972], relevance weighting methods [Robertson and Jones, 1976], and TF-IDF and its variations [Salton and Buckley, 1988].

To measure the semantic similarity of pairs of words several statistical measures have been used. For instance, chi-square statistics [Makoto et al., 1976], pair-wise mutual information [Church and Hanks, 1990], Dice coefficient [Smadja, 1993], log-likelihood ratio [Dunning, 1993], and Jaccard similarity measure [Grefenstette, 1994].

In the last two decades since the seminal papers of [Deerwester et al., 1990, Landauer and Dumais, 1997, Schütze, 1998] Latent Semantic Analysis (LSA) have proven to be useful in several NLP tasks. In LSA a rank-reduction technique is performed on a term document matrix to correlate semantically related terms that are latent in a collection of documents. Amongst many others, they have been applied to solving the TOEFL synonym test [Landauer and Dumais, 1997, Rapp, 2004], automatic thesaurus construction [Schütze, 1998], identification of translation equivalents [Rapp, 1999], word sense induction and discrimination [Schütze, 1998], POS induction [Schütze, 1995], identification of analogical relations [Turney, 2006], PP attachment disambiguation [Pantel and Lin, 2000], and semantic classification [Versley, 2008].

## 2.5 Combining Knowledge-based and Corpus-based similarity

Although the vocabulary of WordNet is very extensive, sometimes we are in the case that a give a word is not included in WordNet (or other dictionary). In these cases it is possible to use other words with similar meaning, or even better, synonyms. It is possible to search in corpora, using distributional semantics, words with similar senses to those words that are not in our dictionary, in order to discover others who are.

[Agirre et al., 2009] explored this approach, improving their results .

## 2.6 Datasets

Datasets for STS are scarce. However, there are at least two important datasets for semantic textual similarity.

The first one, RG, consists of 65 pairs of words collected by [Rubenstein and Goodenough, 1965], who had them judged by 51 human subjects in a scale from 0.0 to 4.0 according to their similarity, but ignoring any other possible semantic relationships that might appear between the terms.

The second dataset, WordSim-353 [Finkelstein et al., 2002] contains 353 word pairs, each associated with an average of 13 to 16 human judgements. In this case, both similarity and relatedness are annotated without any distinction. Several studies indicate that the human scores consistently have very high correlations with each other [Miller et al., 1991, Resnik, 1995], thus validating the use of these datasets for evaluating semantic similarity.

### 2.6.1 RG dataset

This dataset is a consequence of a study about the relationship between similarity of context and similarity of meaning (synonymy). Rubenstein and Goodenough asked to humans how the proportion of words common to context containing a word  $A$  and to the contexts containing a word  $B$  was related to the degree to which  $A$  and  $B$  were similar in meaning. These method assume that pairs of words which have many contexts in common are semantically closely related.

Using 65 pairs of words (which range from highly synonymous pairs to semantically unrelated pairs) the relation is shown between similarity of meaning The 65 word pairs consist of ordinary English words.

#### 2.6.1.1 Procedure

Each subject was given a shuffled deck of 65 slips of paper, each slip containing a different theme pair, and the following instructions:

1. After looking through the whole deck, order the pairs according to amount of 'similarity of meaning' so that the slip containing the pair exhibiting the greatest amount of 'similarity of meaning' is at the top of the deck and the pair exhibiting the least amount is on bottom.
2. Assign a value from 4.0-0.0 to each pair (the greater the 'similarity of meaning', the higher the number). You may assign the same value to more than one pair.

Two groups of college undergraduates were paid to serve as subjects. *Group I*, consisting of 15 subjects, met for two sessions two weeks separated. In the first session the gave synonymy judgments on 48 pairs of themes including 36 of the pairs finally selected for the study. In the second session they gave synonymy judgments on the 65 pairs finally selected. Thus there were 36 theme pair used in both sessions. These pairs enabled Rubenstein and Goodenough to compute the intra-subject reliability in judging synonymy. The product-moment correlation was computed between the first and second judgments on these 36 pairs for each subject. The average correlation over all 15 subject was **.58**.

A second group of 36 subjects (*Group II*) participated only in the second session, on all 65 pairs. A mean judgment was calculated for *Group I* and *II* independently. The correlation between the two group was **.99**. The final synonymy values collected in the dataset are the means of the judgments collected at the second experimental session from both groups, totaling 51 subjects.

### 2.6.2 WordSim-353 dataset

The WordSimilarity-353 Test Collection contains two sets of English word pairs along with human-assigned similarity judgements. The collection can

be used to train and/or test computer algorithms implementing semantic similarity measures.

The first set (set1) contains 153 word pairs along with their similarity scores assigned by 13 subjects. The second set (set2) contains 200 word pairs, with their similarity assessed by 16 subjects. Subjects' names have been replaced by ordinal numbers (1..13, or 1..16) to protect their privacy; identical numbers in the two sets do not necessarily correspond to the same individual.

Each set provides the raw scores assigned by each subject, as well as the mean score for each word pair. For convenience, there is a combined set (combined) that contains a list of all 353 words, along with their mean similarity scores. The combined set is merely a concatenation of the two smaller sets.

[Agirre et al., 2009] also proposed to split the WordSimilarity-353 collection into two datasets, one focused on measuring similarity, and the other one on relatedness.

### 2.6.2.1 Procedure

All the subjects in both experiments possessed near-native command of English. Their instructions were to estimate the relatedness of the words in pairs on a scale from 0 (totally unrelated words) to 10 (very much related or identical words). Specifically, the instructions given to the subjects were the following:

1. Please fill in the similarity scores in the appropriate column of the table. To facilitate processing your questionnaire, please do not print the document but rather type in the values in the table provided.
2. If you do not know the meaning of a particular word - please use a dictionary, or ask a native English speaker.
3. Please DO NOT consult your friends on assigning the similarity scores. It is highly important that the scores you assign be independent of someone else assessment.
4. When estimating similarity of antonyms, consider them 'similar' (i.e., belonging to the same domain or representing features of the same concept), rather than 'dissimilar'.

### 2.6.3 Li and Lee datasets

Other existing datasets are [Li et al., 2006] and [Lee et al., 2005]. The first dataset includes 65 sentence pairs which correspond to the dictionary definitions for the 65 word pairs in Similarity [Rubenstein and Goodenough, 1965]. The authors asked human informants to assess the meaning of the sentence pairs on a scale from 0.0 (minimum similarity) to 4.0 (maximum

similarity). While the dataset is very relevant to STS, it is too small to train, develop and test typical machine learning based systems.

The second dataset comprises 50 documents on news, ranging from 51 to 126 words. Subjects were asked to judge the similarity of document pairs on a five-point scale (with 1.0 indicating “highly unrelated” and 5.0 indicating “highly related”). This second dataset comprises a larger number of document pairs, but it goes beyond sentence similarity into textual similarity.

## 2.7 Conclusions

This chapter has reviewed the state of the art in the area of semantic textual similarity. The concepts of similarity and relatedness have been defined, and we also presented several methods and datasets used for computing semantic similarity and relatedness between textual items.

## Chapter 3

# Applying similarity within PATHS project

Personalized access to cultural heritage spaces (PATHS) is an European project which aims to make it easy to users the exploration of cultural heritage material, suggesting and guiding them through paths.

This chapter collects the work done within the project PATHS to compute the similarity between the items of this cultural heritage.

### 3.1 Introduction

In the previous chapter we saw different methods to measure semantic similarity. In this chapter we describe practical use of one of those methods applying it in a manually created dataset. This dataset was extracted from Europeana (see Section 3.2), in particular from two collections: Culture Grid and SCRAN. The accuracy of our method has been measures using a gold-standard created using human judgments on the dataset.

### 3.2 Europeana

Europeana is a internet portal that acts as an interface for large collections containing millions of books, paintings, films and other museum objects that have been digitized throughout Europe. It is a resource that stores the cultural heritage of Europe. It was funded by the European Commission under its eContentplus programme, one of the research and development funding streams of i2010.

Europeana gives access to different types of content from different types of heritage institutions by querying or browsing the collections. The digital objects that users can find in Europeana are not stored on a central computer, but remain with the cultural institution and hosted on their networks. Europeana collects metadata for each item including a small image. The metadata stores information about the title, the collection, the year of creation, subject, description and more for each item. In order to make the

information searchable, it has to be mapped to a single common standard, known as the Europeana Semantic Elements (ESE). The ESE is an XML Schema that contains the collection identifier, the europeana URI, and the title, creator, description, source, date, year and more.

In Figure 3.1 is an example of an ESE. Figure 3.2 shows the an item as shown on Europeana website.

```
<record>
  <dc:identifier>http://www.picturethepast.org.uk/frontend.php?keywords=Ref_No_increment;EQUALS;NCCW001197</dc:identifier>
  <europeana:uri>http://www.europeana.eu/resolve/record/09405/C052AA1727D9C258801CF676473953A0861A47C0</europeana:uri>
  <dc:title>The Major Oak</dc:title>
  <dc:source>Picture the Past OAI feed</dc:source>
  <dc:contributor>North East Midland Photographic Record</dc:contributor>
  <dc:description>The largest Oak tree in England, perhaps in the world, this famous tree has withstood lightning,
    the drying-out of its roots and even a recent fire. The hollow tree has a circumference of 32 feet
    and the spread of its branches makes a ring 260 feet round.</dc:description>
  <dcterms:isPartOf>Picture the Past</dcterms:isPartOf>
  <dc:language>EN-GB</dc:language>
  <dc:publisher>North East Midland Photographic Record</dc:publisher>
  <dc:subject>Robin_Hood</dc:subject>
  <dc:type>Image</dc:type>
  <dc:format>JPEG/IMAGE</dc:format>
  <europeana:provider>CultureGrid</europeana:provider>
  <europeana:hasObject>true</europeana:hasObject>
  <europeana:country>uk</europeana:country>
  <europeana:type>IMAGE</europeana:type>
  <europeana:language>en</europeana:language>
</record>
```

Figure 3.1: Example of a Europeana Semantic Element

### 3.3 Creating a dataset

The items chosen for the dataset were extracted from two collections (Culture Grid and SCRAN) and stored in three XML files. Culture Grid<sup>1</sup> is a collection of artworks from United Kingdom. SCRAN<sup>2</sup> and the second is . Scran is a online learning resource which contains images of museum exhibits from Scotland.

The Scran collection is stored in a single XML file (00401\_Ag\_UK\_Scran.-oai\_scran.xml) which contains 310802 items. The Culture Grid is stored in two files (09405\_Ag\_UK\_ELocal.xml, 09405a\_Ag\_UK\_ELocal.xml) which contain 381449 and 93105 items respectively.

Our partners from the University of Sheffield selected 30 pairs of items randomly by extracting and storing the URIs of each pair. The final data set consists of 18 pairs from Culture Grid and 12 pairs from Scran. Table 3.1 shows the distribution of items and pairs in each collection.

Dataset	Number of items	Pairs
09405_Ag_UK_ELocal.xml (Culture Grid)	381449 (48.6%)	14
09405a_Ag_UK_ELocal.xml (Culture Grid)	93105 (11.9%)	4
00401_Ag_UK_Scran_oai_scran.xml (Scran)	310802 (39.5%)	12
Total	785356 (100%)	30

Table 3.1: Number of items and selected pairs in each XML file

<sup>1</sup><http://www.culturegrid.org.uk>

<sup>2</sup><http://www.scran.ac.uk>



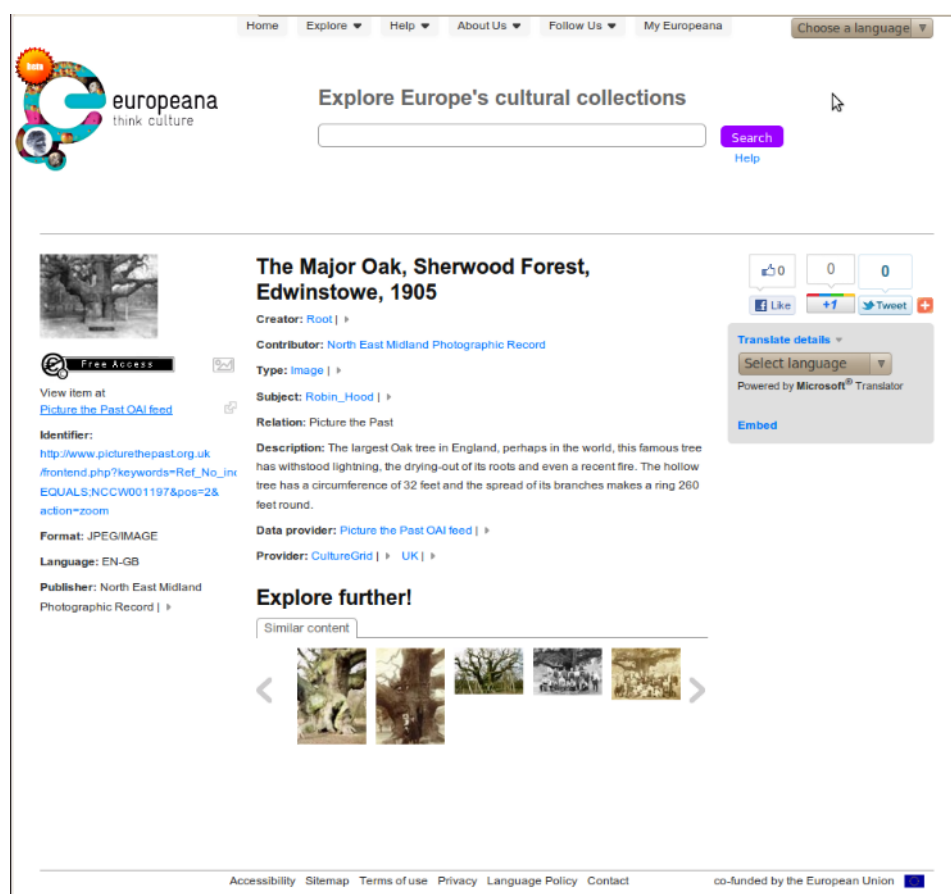


Figure 3.2: Europeana item



The basic criterion for the pair selection is distance in the XML files (number of records between them). There set five different categories of distances: 3 or less, between 3 and 10, between 10 and 30, between 30 and 100 and 100 or more. This distribution is shown in Table 3.2 where  $D$  is the distance between two items. The relative size of each collection was used to select pairs in order to have equal number of pairs in each distance category.

	$D < 3$	$3 < D \leq 10$	$10 < D \leq 30$	$30 < D \leq 100$	$100 > D$
09405	2	3	3	3	3
09405a	2	3	3	3	3
00401	2	3	3	3	3

Table 3.2: Number of pairs from each file in each distance category

In order to calculate distance between items, a Python script was created, which matches all items' URIs using a regular expression in each file. Then, all the URIs were stored in a text file (one per line). Finally, 30 random URIs in each text file were selected and their distance in line numbers was

The interface displays two items side-by-side for comparison. Each item has a photograph, a title, and a block of metadata including title, date, creator, description, and language.

Item 1: Ice house, Keiss	Item 2: Hand operated crab winch, Brora
	
<b>Title:</b> Ice house, Keiss	<b>Title:</b> Hand operated crab winch, Brora
<b>Date:</b> 2002-01-01 00:00:00 ; 2002-12-31 00:00:00 ; October 2002 taken	<b>Date:</b> 2002-01-01 00:00:00 ; 2002-12-31 00:00:00 ; October 2002 taken
<b>Creator:</b> Alastair R. Walker photographer George Dempster, Scottish merchant mentioned ; SECF Project	<b>Creator:</b> Alastair R. Walker photographer ; SECF Project
<b>Description:</b> Mounted 35 mm colour transparency; in good condition; showing Keiss ice house built into a grassy bank. Scotland, Highland, Keiss locality	<b>Description:</b> Mounted 35 mm colour transparency; in good condition; showing a hand operated crab winch with two handles beside a stone wall. Scotland, Highland, Brora locality Scotland, Highland, River Brora mentioned
<b>Language:</b> en	<b>Language:</b> en

**Choose one of the following answers**

- 4 - Very Related/Identical
- 3 - Related
- 2 - Partly Related
- 1 - Unrelated
- 0 - Completely Unrelated

Figure 3.3: Interface used to obtain human judgments

computed by subtracting their positions (using the absolute value).

### 3.3.1 Human judgments on the dataset

In order to create a gold-standard with the selected pairs it was necessary to obtain human judgments of similarity. Human judgments are usually used for creating a gold-standard measure of relatedness between cultural heritage items in Europeana collections. Those data were collected through an on-line survey during a month period. Participants were presented with pairs of items, including images and additional textual information, and asked to judge how similar they are on a scale from 0-4. If two items have a similarity value of 0 means that they have no relations, and a value of 4 means that they are completely related. All the subjects were asked to judge all of the 30 pairs selected in Section 3.3. In Figure 3.3 we can see the interface if the survey.

A total of 74 evaluations were received, with 38 of them completely completed, but 36 were incompleted. For this experiment, only the completed ones are used. The average of the responses for each pair was calculated to create the gold-standard. The highest average for a pair was 4 (Pair 15) and the lowest was 0.66 (Pair 20).

### 3.4 Computing similarity using random walks

The semantic disambiguation UKB<sup>3</sup> algorithm [Agirre and Soroa, 2009b] applies personalized PageRank on a graph generated from the English WordNet. This algorithm has proven to be very competitive and it is easily portable to other languages that have a wordnet, with good results [Agirre et al., 2010]. But it also has other utilities, and can be used to propagate information through the WordNet structure.

To compute similarity using UKB we represent WordNet as a graph  $G = (V, E)$  as follows: graph nodes represent WordNet concepts (synsets) and dictionary words; relations among synsets are represented by undirected edges; and dictionary words are linked to the synsets associated to them by directed edges.

Given a pair of words (or vectors of words) and a graph-based representation of WordNet, our method has basically two steps: We first compute the personalized PageRank over WordNet separately for each of the words, producing a probability distribution over WordNet synsets. We then compare how similar these two discrete probability distributions are by encoding them as vectors and computing the cosine between the vectors. We present each step in turn.

Once personalized PageRank is computed, it returns a probability distribution over WordNet synsets. The similarity between two words can thus be implemented as the similarity between the probability distributions. Alternatively, we can interpret the probability distribution for a word  $w$  as a vector  $w$  of weights  $w_i$  where each dimension  $i$  is a synset, and use the cosine to compute similarity.

The similarity algorithm in the UKB package requires the introduction of the words in a particular format. This means that for every word introduced there must be a lemma and part-of-speech tagged. For example, suppose we want to calculate the similarity of the following two sentences:

- Someone is greating a carrot.
- A woman is grating an orange food.

The first thing we must do is to lemmatize the sentences and get the lemma and the part-of-speech (PoS) of every word in the sentence. We used the Stanford parser which returns the sentences with the following format: *word/part-of-speech/entity-type/lemma*. For the previous strings, we obtain this analysis (lemmas and PoS are marked in bold for better visualization):

- Someone is greating a carrot.  
Someone-0/**NN**/O/**someone** is-1/**VBZ**/O/**be** greating-2/**VBG**/O/-  
/**great** a-3/**DT**/O/**a** carrot-4/**NN**/O/**carrot** .-5/./O/.

---

<sup>3</sup><http://ixa2.si.ehu.es/ukb/>

- A woman is grating an orange food.  
A-0/DT/O/a woman-1/NN/O/woman is-2/VBZ/O/be grating-3/-  
/VBG/O/grate an-4/DT/O/a orange-5/JJ/O/orange food-6/NN/  
/O/food .-7/./O/.

Once we have the output of the parser we can construct the input to the UKB similarity algorithm. UKB makes use of WordNet as knowledge base, that is why UKB is only capable of using nouns, verbs, adjectives and adverbs. For the above example the input would be as follows:

- someone#n#w1#1 is#v#w2#1 great#v#w3#1 carrot#n#w4#1
- woman#n#w1#1 is#v#w2#1 grate#v#w3#1 orange#a#w4#1  
food#n#w5#1

Initializing the respective random walks with the those lemmas in each sentence in turn, we obtain the two vectors of synsets, and using the cosine or dot product, it produces the similarity value.

### 3.4.1 Experiment setup

The process of the experiment is divided into four steps:

1. After obtaining the lemma and PoS a stop-word-list is used to remove the stop-words, which introduce noise to the algorithm.
2. All those words that are not nouns, verbs, adjectives or adverbs are removed.
3. Similarity algorithm from UKB package is called to get the similarity scores between each pair of items.
4. Finally, the correlation between the results obtained by our system and the results of the gold standard is computed using Spearman<sup>4</sup> and Pearson<sup>5</sup> correlation measures.

Following this procedure different experiments were carried out by changing some parameters: different graphs for UKB, selecting different features from items and computing the similarity value using cosine or dot product (see Section 3.5).

## 3.5 Results

To generate the results of this section several parameters were employed.

Four different knowledge bases have been generated:

<sup>4</sup>[http://en.wikipedia.org/wiki/Spearman%27s\\_rank\\_correlation\\_coefficient](http://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient)

<sup>5</sup>[http://en.wikipedia.org/wiki/Pearson\\_product-moment\\_correlation\\_coefficient](http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient)

1. **wnet30**: Original WordNet relations.
2. **wnet30g**: Also contains the relationships between glosses, increasing the size and richness of the knowledge base.
3. **wnet30gk1**: Adds KnowNet-5 to the second knowledge base. KnowNet (KN)[Cuadros and Rigau, 2008] is an extensible, large and accurate knowledge base, which has been derived by semantically disambiguating small portions of Topic Signatures acquired from the web. KnowNet-5 is obtained by disambiguating only the first five words from each Topic Signature.
4. **pre(wnet30g.1000)**: Instead of generating vectors using PageRank, precomputed vectors with the 1000 most relevant nodes are used. This method is faster, and is expected to improve results, as they can reduce the noise.

Instructions for preparing the binary databases for UKB using WordNet relations are inside the downloadable file<sup>6</sup> of the UKB package.

The UKB similarity algorithm can calculate the similarity by two different ways:

1. **cos**: Using the cosine.
2. **dot**: Using the dot or scalar product.

The metadata of Europeana items includes information about the title, the digital format, the collection, the year of creation, a small description and other features of each item (see Section 3.2). For these experiments the following sets of features will be used:

1. **Title**: Titles of the items.
2. **Title+subject**: Titles and subjects (keywords) of the items.
3. **Title+description**: Titles and descriptions of the items.
4. **Title+subject+description**: Titles, subjects and descriptions of the items.

It is important to note that some items have no subject or description.

---

<sup>6</sup><http://ixa2.si.ehu.es/ukb/>

### 3.5.1 Using WordNet 3.0

Results for Culture Grid are shown in Table 3.3 and the results for SCRAN in Table 3.4. The highest accuracy is shown in bold and the lowest one is shown in italics.

The highest values for Culture Grid are obtained by computing the cosine similarity (.70 for Spearman and .77 for and Pearson). The highest values for SCRAN are obtained using dot (.63 for Spearman) and cos (.66 to Pearson).

Regarding to the lower values, in the case of Spearman they are obtained with cos (.64 for Culture Grid and .48 for SCRAN), and for Pearson with dot (.39 for Culture Grid and .56 for SCRAN).

Culture Grid			
Sim	Features	Spearman	Pearson
cos	Title	<b>.700893</b>	.766065
	Title+subject	.661142	<b>.772808</b>
	Title+description	<i>.649123</i>	.726740
	Title+subject+description	.667699	.737717
dot	Title	.672967	.414827
	Title+subject	.682852	<i>.393197</i>
	Title+description	.675955	.674189
	Title+subject+description	.684211	.673117

Table 3.3: Results for Culture Grid using WordNet 3.0 as KB

SCRAN			
Sim	Features	Spearman	Pearson
cos	Title	.630849	.668852
	Title+subject	.598950	<b>.669579</b>
	Title+description	<i>.489511</i>	.617608
	Title+subject+description	.503497	.635467
dot	Title	.630849	.612369
	Title+subject	<b>.633977</b>	.624827
	Title+description	.510490	<i>.568123</i>
	Title+subject+description	.503497	.574064

Table 3.4: Results for SCRAN using WordNet 3.0 as KB

### 3.5.2 Using WordNet 3.0 enriched with gloss relations

Results for Culture Grid are shown in Table 3.5 and the results for SCRAN in Table 3.6. The highest accuracy is shown in bold and the lowest one is shown in italics.

The highest values for Culture Grid are obtained by computing the cosine similarity (.75 for Spearman and .77 for Pearson). The highest values for SCRAN are obtained using cos (.68 for Spearman and .67 for Pearson), although in the case of Spearman the best result obtained with cos is equal to the best results obtained with dot. These maxims represent an improvement regarding to wnet30 of about 2%-5% (Spearman-Pearson) for Culture Grid and 5%-2% for SCRAN.

Regarding to the lower values, in the case of Spearman they are obtained with cos (.64 for Culture Grid and .50 for SCRAN), and for Pearson with dot (.53 for Culture Grid and .54 for SCRAN). The differences are not significant except for Pearson and dot similarity, which presents an improvement of 14% in the SCRAN collection.

Given these results it seems that the glosses provide valuable information to the UKB similarity algorithm.

Culture Grid			
Sim	Features	Spearman	Pearson
cos	Title	<b>.755700</b>	.768064
	Title+subject	.659066	<b>.773989</b>
	Title+description	<i>.649123</i>	.696836
	Title+subject+description	.665635	.708016
dot	Title	.695921	.553593
	Title+subject	.722108	<i>.539430</i>
	Title+description	.737874	.631507
	Title+subject+description	.737874	.629734

Table 3.5: Results for Culture Grid using wnet30g as KB

SCRAN			
Sim	Features	Spearman	Pearson
cos	Title	<b>.681607</b>	<b>.671476</b>
	Title+subject	<i>.507882</i>	.669528
	Title+description	.531469	.606218
	Title+subject+description	.545454	.627766
dot	Title	<b>.681607</b>	.631844
	Title+subject	.549913	.624880
	Title+description	.566434	<i>.540513</i>
	Title+subject+description	.566434	.544847

Table 3.6: Results for SCRAN using wnet30g as KB

### 3.5.3 Using WordNet 3.0 enriched with gloss relations and KnowNet-5

Results for Culture Grid are shown in Table 3.7 and the results for SCRAN in Table 3.8. The highest accuracy is shown in bold and the lowest one is shown in italics.

The highest values for Culture Grid are obtained by computing the cosine similarity (.76 for Spearman and .77 for Pearson). The highest value for SCRAN are obtained using cos (.71 for Spearman and .67 for Pearson), although in the case of Spearman the best result obtained with cos is equal to the best results obtained with dot. These maxims represent an improvement regarding to wnet30g of about 1%-0% (Spearman-Pearson) for Culture Grid and 3%-0% for SCRAN.

Regarding to the lower values, in the case of Spearman they are obtained with cos (.51 for Culture Grid) and with dot (.51 to SCRAN), and for Pearson with dot (.56 for Culture Grid and .54 for SCRAN). In this case, the differences are not very significant except for the case using Spearman and cos, having a drop of a 13% in the Culture Grid collection.

Given these results is not possible to decide whether KnowNet-5 incorporates some improvement or not. For Pearson hardly changed the results. For Spearman, maximum values have increased, but in turn, the minimum values have decreased.

Culture Grid			
Sim	Features	Spearman	Pearson
cos	Title	<b>.762024</b>	.767849
	Title+subject	<i>.613399</i>	<b>.773954</b>
	Title+description	.649123	.698066
	Title+subject+description	.669763	.708993
dot	Title	.714701	.570577
	Title+subject	.701447	<i>.562844</i>
	Title+description	.715170	.626460
	Title+subject+description	.715170	.623881

Table 3.7: Results for Culture Grid using wnet30gk1 as KB

### 3.5.4 Using pre-calculated vectors from WordNet 3.0 enriched with gloss relations

Results for Culture Grid are shown in Table 3.9 and the results for SCRAN in Table 3.10. The highest accuracy is shown in bold and the lowest one is shown in italics.

The highest values for Culture Grid are obtained by computing the cosine similarity (.69 for Spearman and .77 for Pearson). The highest values for SCRAN are obtained using dot (.65 for Spearman) and cos (.67 for Pearson).



SCRAN			
Sim	Features	Spearman	Pearson
cos	Title	<b>.717863</b>	<b>.671172</b>
	Title+subject	.423818	.669641
	Title+description	.531469	.618808
	Title+subject+description	.552448	.637992
dot	Title	<b>.717863</b>	.631076
	Title+subject	.521892	.623291
	Title+description	<i>.517482</i>	<i>.541128</i>
	Title+subject+description	<i>.517482</i>	.543358

Table 3.8: Results for SCRAN using wnet30gk1 as KB

These maxims do not present any significant change for Pearson, but in the case of Spearman the results are worse than with wnetgk1 (7% for Culture Grid and 6% for SCRAN).

The lower values, in all cases, have been obtained with dot, reaching even close to zero negative values.

These results do not show the expected improvement, but using features like Title or combinations of features like Title-Subject similar results can be achieved saving an enormous amount of calculation time.

Culture Grid			
Sim	Features	Spearman	Pearson
cos	Title	<b>.692461</b>	.770485
	Title+subject	.623778	<b>.777653</b>
	Title+description	.665635	.712608
	Title+subject+description	.644995	.712931
dot	Title	.656273	.504885
	Title+subject	.600207	.562849
	Title+description	<i>-.159959</i>	.005119
	Title+subject+description	<i>-.141383</i>	<i>-.004264</i>

Table 3.9: Results for Culture Grid using pre-calculated vectors from wnet30g.

### 3.5.5 Comparison of the obtained results

With the aim of displaying more clearly the results four types of plots have been generated (four for each collection, therefore eight in total).

1. **Comparison of knowledge bases:** The averages of the results obtained with each of the four knowledge bases employed have been calculated (Figure 3.4 for Culture Grid and Figure 3.5 for SCRAN).

SCRAN			
Sim	Features	Spearman	Pearson
cos	Title	.601844	<b>.671612</b>
	Title+subject	.416813	.664093
	Title+description	.517482	.576186
	Title+subject+description	.517482	.587264
dot	Title	<b>.652602</b>	.340069
	Title+subject	.430824	.456770
	Title+description	<i>-.055944</i>	<i>-.114318</i>
	Title+subject+description	.013986	-.100632

Table 3.10: Results for SCRAN using pre-calculated vectors from wnet30g.

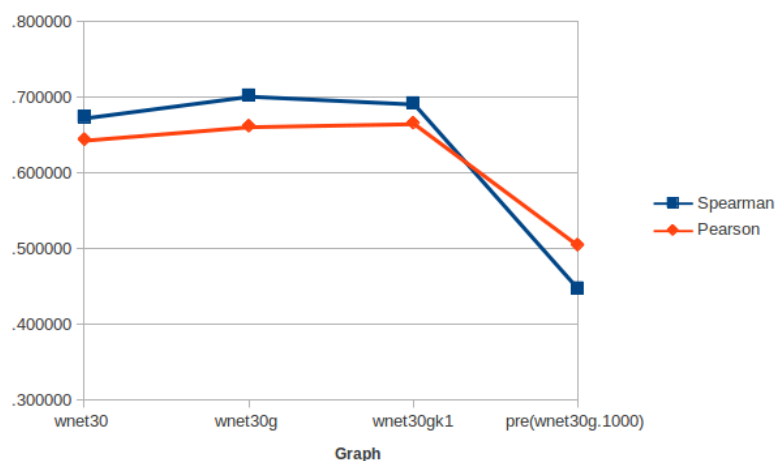


Figure 3.4: Culture Grid: Averages of the results with different knowledge bases.

2. **Comparison of similarity measures:** The averages of the results obtained with the two modes of computing the similarity with UKB (cos and dot) have been calculated (Figure 3.6 for Culture Grid and Figure 3.7 for SCRAN).
3. **Comparison of features:** The averages of the results obtained with each of the features combinations extracted of Europeana metadates features have been calculated (Figure 3.8 for Culture Grid and Figure 3.9 for SCRAN).
4. **Comparison of correlation values:** The maximum, mean and minimum obtained by Spearman and Pearson correlations have been calculated (Figure 3.10 for Culture Grid and Figure 3.11 for SCRAN).

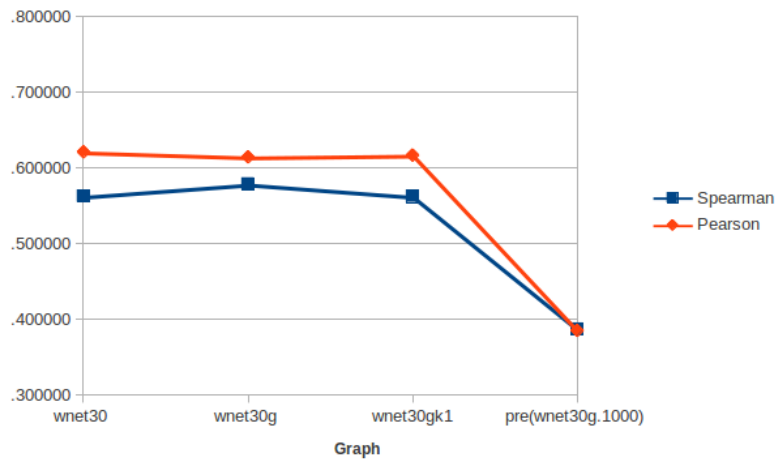


Figure 3.5: SCRAN: Averages of the results with different knowledge bases.

With the assistance of these plots it is possible to visualize better the results, allowing us to draw some conclusions:

- Wnetgk1 and wnetg knowledge bases offer the best results in both collections.
- Pre-computed vectors obtained, by far, the worst results in both collections.
- The correlations obtained using the cosine are better than with the scalar product.
- The similarity calculation using the cosine has greater positive effect on the Pearson correlation.
- On average, the best feature to achieve a better correlation is the Title.

### 3.5.6 Comparison with the state of the art

[Aletras, 2011] created the gold-standard used in this master thesis. They applied different techniques to the dates extracted from items in Europeana and evaluated them against the gold-standard. They used knowledge-based and corpus-based techniques. In this section, we compare the results with that ones obtained by [Aletras, 2011] using techniques based on wordnets. In particular, they used the Extended Lesk Measure.

The results of Table 3.11 show that our technique obtains slightly better results; we obtain .750 with pre-calculated vectors, using Title as features, and they obtain .703 with *Extended Lesk Measure*, using Title and Description as features. [Aletras, 2011] did not discriminate by collections so both Culture Grid and SCRAN are used in conjunction in this experiment. The highest accuracy is shown in bold and the lowest one is shown in italics.

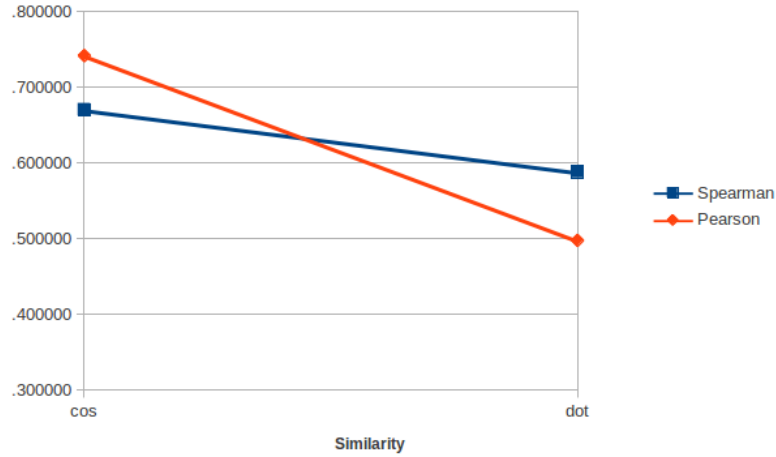


Figure 3.6: Culture Grid: Averages of the results with UKB's similarity modes (cos and dot).

Measure	Features		
	Title	Title+subject	Title+description
wnet30 (cos)	.745	.631	.518
wnet30g (cos)	.740	.644	.494
wnet30gk1 (cos)	.740	.644	-
pre(wnet30g.1000)(cos)	<b>.750</b>	.673	.547
Extended Lesk Measure	.557	.400	.703

Table 3.11: Results for SCRAN using pre-calculated vectors from wnet30g.

The best result of this dataset obtained by [Aletras, 2011] was a correlation of .856, using Title as a feature and a technique based on overlap called *Normalized Simple Overlap*.

### 3.6 Conclusions

In this section we draw some conclusions about the work done in this chapter:

- We applied a graph-based method that applies personalized PageRank on graphs generated from wordnet to compute the similarity between elements of Europeana.
- We tested different knowledge-bases for the personalized PageRank algorithm.
- We tried different ways to compute the similarity with the UKB package: *cosine* and *dot product*.

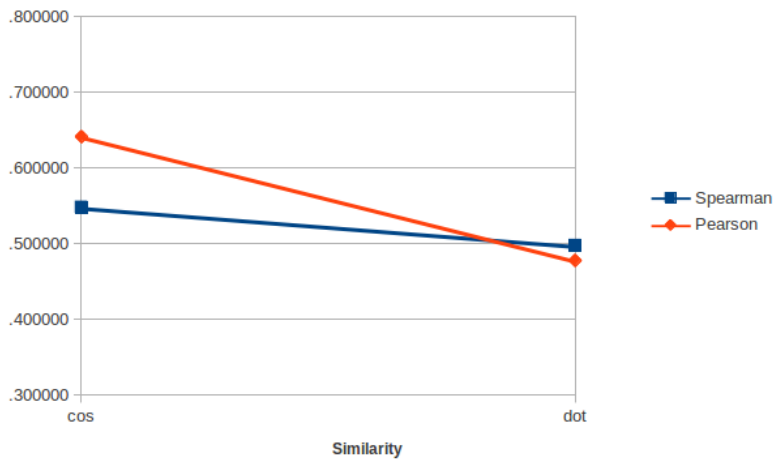


Figure 3.7: SCRAN: Averages of the results with UKB’s similarity modes (cos and dot).

- We tested with different combinations of features extracted from Europeana elements.
- We compared between two different methods to calculate the correlation between two similarity results: *Spearman* and *Pearson*.

After this experiment, we can say that the combination that works better for this method is the one with the following configuration: pre-calculated vectors (1000 most relevant nodes) from WordNet enriched with glosses, similarity computed with *cosine*, correlation computed with *Pearson* and with *Title+Subject* as features on the Europeana item. These results are better than those obtained by previous techniques that use wordnets.

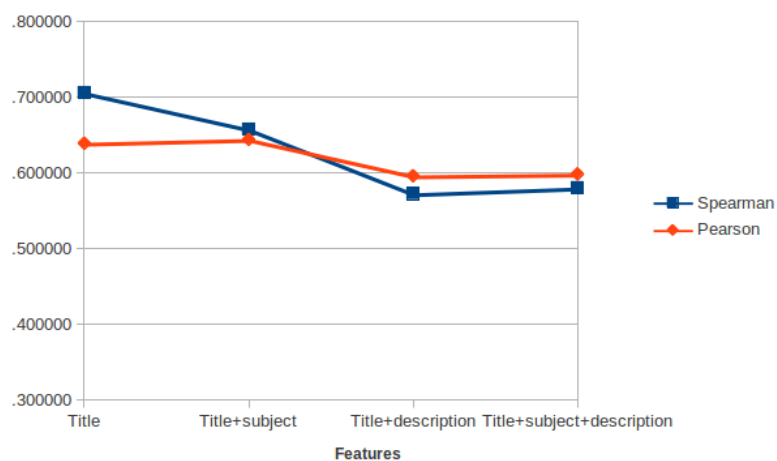


Figure 3.8: Culture Grid: Averages of the results obtained with the different features combinations.

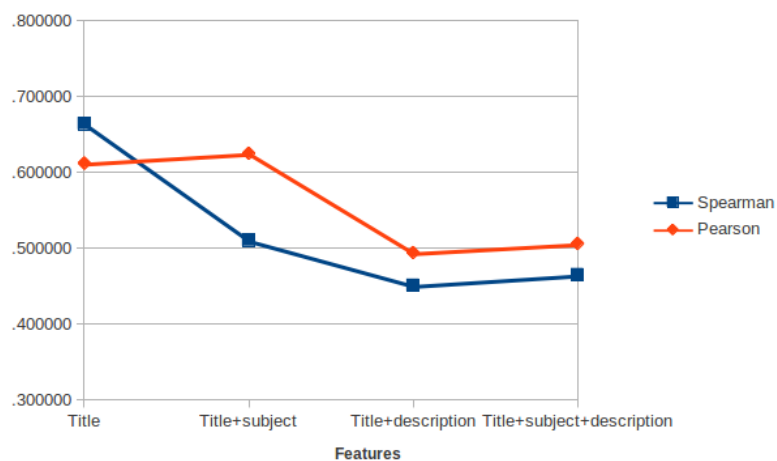


Figure 3.9: SCRAN: Averages of the results obtained with the different features combinations.

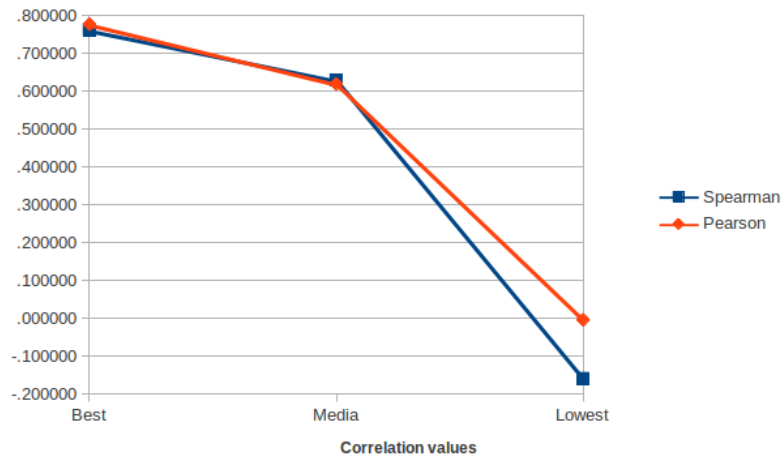


Figure 3.10: Culture Grid: Maximum, mean and minimum obtained by Spearman and Pearson correlations.

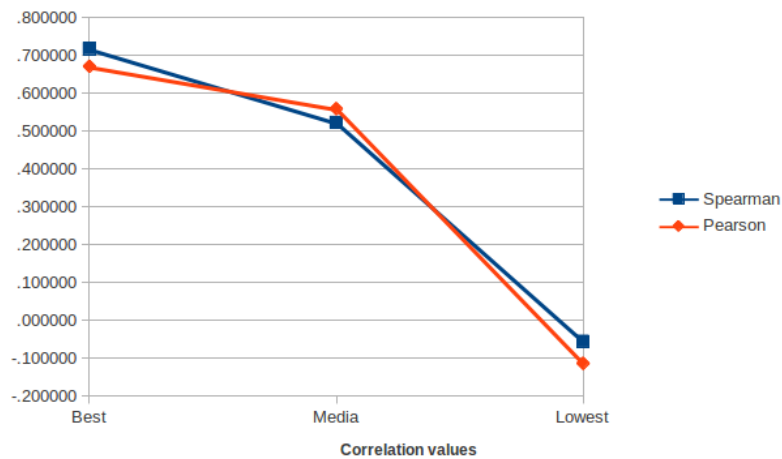


Figure 3.11: SCRAN: Maximum, mean and minimum obtained by Spearman and Pearson correlations.





## Chapter 4

# SemEval-2012 Task 6

This chapter describes the work done for the Semantic Textual Similarity (STS) shared task of SemEval-2012. This work has involved supporting the creation of datasets for similarity tasks, as well as the organization of the task itself. Although it was not the primary objective of this thesis, a system was also presented to the competition.

### 4.1 Introduction

Semantic Textual Similarity (STS) measures the degree of semantic equivalence between two sentences. STS is related to both Textual Entailment (TE) and Paraphrase (PARA). STS is more directly applicable in a number of NLP tasks than TE and PARA such as Machine Translation and evaluation, Summarization, Machine Reading, Deep Question Answering, etc.

STS differs from TE in as much as it assumes symmetric graded equivalence between the pair of textual snippets. In the case of TE the equivalence is directional, e.g. a car is a vehicle, but a vehicle is not necessarily a car. Additionally, STS differs from both TE and PARA in that, rather than being a binary yes/no decision (e.g. a vehicle is not a car), STS incorporates the notion of graded semantic similarity (e.g. a vehicle and a car are more similar than a wave and a car).

STS provides a unified framework that allows for an extrinsic evaluation of multiple semantic components that otherwise have tended to be evaluated independently and without broad characterization of their impact on NLP applications. Such components include word sense disambiguation and induction, lexical substitution, semantic role labeling, multiword expression detection and handling, anaphora and coreference resolution, time and date resolution, named-entity handling, underspecification, hedging, semantic scoping and discourse analysis. Though not in the scope of the current pilot task, we plan to explore building an open source toolkit for integrating and applying diverse linguistic analysis modules to the STS task.

While the characterization of STS is still preliminary, we observed that

there was no comparable existing dataset extensively annotated for pairwise semantic sentence similarity. We approached the construction of the first STS dataset with the following goals: (1) To set a definition of STS as a graded notion which can be easily communicated to non-expert annotators beyond the likert-scale; (2) To gather a substantial amount of sentence pairs from diverse datasets, and to annotate them with high quality; (3) To explore evaluation measures for STS; (4) To explore the relation of STS to PARA and Machine Translation Evaluation exercises.

In the next section we present the various sources of the STS data and the annotation procedure used. Section 4.4 investigates the evaluation of STS systems. Section 4.6 summarizes the resources and tools used by participant systems. Finally, Section 4.9 draws some conclusions.

## 4.2 Source Datasets

When constructing our datasets, gathering naturally occurring pairs of sentences with different degrees of semantic equivalence was a challenge in itself. If we took pairs of sentences at random, the vast majority of them would be totally unrelated, and only a very small fragment would show some sort of semantic equivalence. Accordingly, we investigated reusing a collection of existing datasets from tasks that are related to STS.

We first studied the pairs of text from the Recognizing TE challenge. The first editions of the challenge included pairs of sentences as the following:

T: The Christian Science Monitor named a US journalist kidnapped in Iraq as freelancer Jill Carroll.  
H: Jill Carroll was abducted in Iraq.

The first sentence is the text, and the second is the hypothesis. The organizers of the challenge annotated several pairs with a binary tag, indicating whether the hypothesis could be entailed from the text. Although these pairs of text are interesting we decided to discard them from this pilot because the length of the hypothesis was typically much shorter than the text, and we did not want to bias the STS task in this respect. We may, however, explore using TE pairs for STS in the future.

Microsoft Research (MSR) has pioneered the acquisition of paraphrases with two manually annotated datasets. The first, called MSR Paraphrase (MSRpar for short) has been widely used to evaluate text similarity algorithms. It contains 5801 pairs of sentences gleaned over a period of 18 months from thousands of news sources on the web [Dolan et al., 2004]. 67% of the pairs were tagged as paraphrases. The inter annotator agreement is between 82% and 84%. Complete meaning equivalence is not required, and the annotation guidelines allowed for some relaxation. The pairs which were annotated as not being paraphrases ranged from completely unrelated semantically, to partially overlapping, to those that were almost-but-not-quite semantically equivalent. In this sense our graded annotations enrich



- A person is slicing a cucumber into pieces.
- A chef is slicing a vegetable.
- A person is slicing a cucumber.
- A woman is slicing vegetables.
- A woman is slicing a cucumber.
- A person is slicing cucumber with a knife.
- A person cuts up a piece of cucumber.
- A man is slicing cucumber.
- A man cutting zucchini.

Figure 4.1: Video and corresponding descriptions from MSRvid

the dataset with more nuanced tags, as we will see in the following section. We followed the original split of 70% for training and 30% for testing. A sample pair from the dataset follows:

The Senate Select Committee on Intelligence is preparing a blistering report on prewar intelligence on Iraq.

American intelligence leading up to the war on Iraq will be criticized by a powerful US Congressional committee due to report soon, officials said today.

In order to construct a dataset which would reflect a uniform distribution of similarity ranges, we sampled the MSRpar dataset at certain ranks of string similarity. We used the implementation readily accessible at CPAN<sup>1</sup> of a well-known metric [Ukkonen, 1985]. We sampled equal numbers of pairs from five bands of similarity in the [0.4 .. 0.8] range separately from the paraphrase and non-paraphrase pairs. We sampled 1500 pairs overall, which we split 50% for training and 50% for testing.

The second dataset from MSR is the MSR Video Paraphrase Corpus (MSRvid for short). The authors showed brief video segments to Annotators from Amazon Mechanical Turk (AMT) and were asked to provide a one-sentence description of the main action or event in the video [Chen and Dolan, 2011]. Nearly 120 thousand sentences were collected for 2000 videos. The sentences can be taken to be roughly parallel descriptions, and they included sentences for many languages. Figure 4.1 shows a video and corresponding descriptions.

The sampling procedure from this dataset is similar to that for MSRpar. We construct two bags of data to draw samples. The first includes all

<sup>1</sup><http://search.cpan.org/~mlehmman/String-Similarity-1.04/Similarity.pm>

## Compare Two Similar Sentences

Score how similar two sentences are to each other according to the following scale.

The sentences are:

- (5) Completely equivalent**, as they *mean the same thing*.
- (4) Mostly equivalent**, but some *unimportant details differ*.
- (3) Roughly equivalent**, but some *important information differs/missing*.
- (2) Not equivalent**, but *share some details*.
- (1) Not equivalent**, but are *on the same topic*.
- (0) On different topics**.

Select a similarity rating for each sentence pair below:

Figure 4.2: Definition and instructions for annotation

possible pairs for the same video, and the second includes pairs taken from different videos.

Note that not all sentences from the same video were equivalent, as some descriptions were contradictory or unrelated. Conversely, not all sentences coming from different videos were necessarily unrelated, as many videos were on similar topics. We took an equal number of samples from each of these two sets, in an attempt to provide a balanced dataset between equivalent and non-equivalent pairs. The sampling was also done according to string similarity, but in four bands in the [0.5 .. 0.8] range, as sentences from the same video had a usually higher string similarity than those in the MSRpar dataset. We sampled 1500 pairs overall, which we split 50% for training and 50% for testing.

Given the strong connection between STS systems and Machine Translation evaluation metrics, we also sampled pairs of segments that had been part of human evaluation exercises. Those pairs included a reference translation and a automatic Machine Translation system submission, as follows:

The only instance in which no tax is levied is when the supplier is in a non-EU country and the recipient is in a Member State of the EU.

The only case for which no tax is still perceived ”is an example of supply in the European Community from a third country.

We selected pairs from the translation shared task of the 2007 and 2008 ACL Workshops on Statistical Machine Translation (WMT) [Callison-Burch et al., 2007, Callison-Burch et al., 2008].

For consistency, we only used French to English system submissions. The training data includes all of the Europarl human ranked fr-en system submissions from WMT 2007, with each machine translation being paired with the correct reference translation. This resulted in 729 unique training pairs.

The test data is comprised of all Europarl human evaluated fr-en pairs from WMT 2008 that contain 16 white space delimited tokens or less.

In addition, we selected two other datasets that were used as out-of-domain testing. One of them comprised of all the human ranked fr-en system submissions from the WMT 2007 news conversation test set, resulting in 351 unique system reference pairs.<sup>2</sup>

The second set is radically different as it comprised 750 pairs of glosses from OntoNotes 4.0 [Hovy et al., 2006] and WordNet 3.1 [Fellbaum, 1998b] senses. The mapping of the senses of both resources comprised 110K sense pairs. The similarity between the sense pairs was generated using simple word overlap. 50% of the pairs were sampled from senses which were deemed as equivalent senses, the rest from senses which did not map to one another.

### 4.3 Annotation

In this first dataset we defined a straightforward likert scale ranging from 5 to 0, but we decided to provide definitions for each value in the scale (cf. Figure 4.2). We first did pilot annotations of 200 pairs selected at random from the three main datasets in the training set. We did the annotation, and the pairwise Pearson ranged from 84% to 87% among ourselves. The agreement of each annotator with the average scores of the other was between 87% and 89%.

In the future, we would like to explore whether the definitions improve the consistency of the tagging with respect to a likert scale without definitions. Note also that in the assessment of the quality and evaluation of the systems performances, we just took the resulting SS scores and their averages. Using the qualitative descriptions for each score in analysis and evaluation is left for future work.

Given the good results of the pilot we decided to deploy the task in Amazon Mechanical Turk (AMT) in order to crowd source the annotation task. The turkers were required to have achieved a 95% of approval rating in their previous HITs, and had to pass a qualification task which included 6 example pairs. Each HIT included 5 pairs of sentences, and was paid at 0.20\$ each. We collected 5 annotations per HIT. In the latest data collection, each HIT required 114.9 second for completion.

In order to ensure the quality, we also performed post-hoc validation. Each HIT contained one pair from our pilot. After the tagging was completed we checked the correlation of each individual turker with our scores, and removed annotations of turkers which had low correlations (below 50%). Given the high quality of the annotations among the turkers, we could alternatively use the correlation between the turkers itself to detect poor quality annotators.

---

<sup>2</sup>At the time of the shared task, this data set contained duplicates resulting in 399 sentence pairs.

## 4.4 Systems Evaluation

Given two sentences,  $s_1$  and  $s_2$ , an STS system would need to return a similarity score. Participants can also provide a confidence score indicating their confidence level for the result returned for each pair, but this confidence is not used for the main results. The output of the systems performance is evaluated using the Pearson product-moment correlation coefficient between the system scores and the human scores, as customary in text similarity [Rubenstein and Goodenough, 1965]. We calculated Pearson for each evaluation dataset separately.

In order to have a single Pearson measure for each system we concatenated the gold standard (and system outputs) for all 5 datasets into a single gold standard file (and single system output).

The first version of the results were published using this method, but the overall score did not correspond well to the individual scores in the datasets, and participants proposed two additional evaluation metrics, both of them based on Pearson correlation. The organizers of the task decided that it was more informative, and on the benefit of the community, to also adopt those evaluation metrics, and the idea of having a single main evaluation metric was dropped. This decision was not without controversy, but the organizers gave more priority to openness and inclusiveness and to the involvement of participants. The final result table thus included three evaluation metrics. For the future we plan to analyze the evaluation metrics, including non-parametric metrics like Spearman.

### 4.4.1 Evaluation metrics

The first evaluation metric is the Pearson correlation for the concatenation of all five datasets, as described above. We will use *overall Pearson* or simply *ALL* to refer to this measure.

The second evaluation metric normalizes the output for each dataset separately, using the linear least squares method. We concatenated the system results for five datasets and then computed a single Pearson correlation. Given  $Y = \{y_i\}$  and  $X = \{x_i\}$  (the gold standard scores and the system scores, respectively), we transform the system scores into  $X' = \{x'_i\}$  in order to minimize the squared error  $\sum_i (y_i - x'_i)^2$ . The linear transformation is given by  $x'_i = x_i * \beta_1 + \beta_2$ , where  $\beta_1$  and  $\beta_2$  are found analytically. We refer to this measure as *Normalized Pearson* or simply *ALLnorm*. This metric was suggested by one of the participants, Sergio Jimenez.

The third evaluation metric is the weighted mean of the Pearson correlations on individual datasets. The Pearson returned for each dataset is weighted according to the number of sentence pairs in that dataset. Given  $r_i$  the five Pearson scores for each dataset, and  $n_i$  the number of pairs in each dataset, the weighted mean is given as  $\sum_{i=1..5} (r_i * n_i) / \sum_{i=1..5} n_i$ . We refer to this measure as *weighted mean of Pearson* or *Mean* for short.

### 4.4.2 Using confidence scores

Participants were allowed to include a confidence score between 1 and 100 for each of their scores. We used weighted Pearson to use those confidence scores<sup>3</sup>. Table 4.2 includes the list of systems which provided a non-uniform confidence. The results show that some systems were able to improve their correlation, showing promise for the usefulness of confidence in applications.

### 4.4.3 The Baseline System

The scores were produced using a simple word overlap baseline system. The input sentences were tokenized splitting at white spaces, and then represented each sentence as a vector in the multidimensional token space. Each dimension had 1 if the token was present in the sentence, 0 otherwise. Similarity of vectors was computed using cosine similarity.

A random baseline was run several times, yielding close to 0 correlations in all datasets, as expected. There are references to the random baseline again in Section 4.5.

### 4.4.4 Participation

Participants could send a maximum of three system runs. After downloading the test datasets, they had a maximum of 120 hours to upload the results. 35 teams participated, submitting 88 system runs (cf. first column of Table 4.1). Due to lack of space we can't detail the full names of authors and institutions that participated. The interested reader can use the name of the runs to find the relevant paper in these proceedings.

There were several issues in the submissions. The submission software did not ensure that the naming conventions were appropriately used, and this caused some submissions to be missed, and in two cases the results were wrongly assigned. Some participants returned Not-a-Number as a score, and the organizers had to request whether those were to be taken as a 0 or as a 5.

Finally, one team submitted past the 120 hour deadline and some teams sent missing files after the deadline. All those are explicitly marked in Table 4.1. The teams that included one of the organizers are also explicitly marked. We want to stress that in these teams the organizers did not allow the developers of the system to access any data or information which was not available for the rest of participants. One exception is *weiwei*, as they generated the 110K OntoNotes-WordNet dataset from which the other organizers sampled the surprise data set.

After the submission deadline expired, the organizers published the gold standard in the task website, in order to ensure a transparent evaluation process.

---

<sup>3</sup>[http://en.wikipedia.org/wiki/Pearson\\_product-moment\\_correlation\\_coefficient#Calculating\\_a\\_weighted\\_correlation](http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient#Calculating_a_weighted_correlation)

## 4.5 Results

Table 4.1 shows the results for each run in alphabetic order.

Each result is followed by the rank of the system according to the given evaluation measure. To the right, the Pearson score for each dataset is given. In boldface, the three best results in each column.

First of all we want to stress that the large majority of the systems are well above the simple baseline, although the baseline would rank 70 on the Mean measure, improving over 19 runs.

The correlation for the non-MT datasets were really high: the highest correlation was obtained was for MSRvid (0.88  $r$ ), followed by MSRpar (0.73  $r$ ) and On-WN (0.73  $r$ ). The results for the MT evaluation data are lower, (0.57  $r$ ) for SMT-eur and (0.61  $r$ ) for SMT-News. The simple token overlap baseline, on the contrary, obtained the highest results for On-WN (0.59  $r$ ), with (0.43  $r$ ) on MSRpar and (0.40  $r$ ) on MSRvid. The results for MT evaluation data are also reversed, with (0.40  $r$ ) for SMT-eur and (0.45  $r$ ) for SMT-News.

The ALLnorm measure yields the highest correlations. This comes at no surprise, as it involves a normalization which transforms the system outputs using the gold standard. In fact, a random baseline which gets Pearson correlations close to 0 in all datasets would attain Pearson of 0.5891<sup>4</sup>.

Although not included in the results table for lack of space, we also performed an analysis of confidence intervals. For instance, the best run according to ALL ( $r = .8239$ ) has a 95% confidence interval of [.8123,.8349] and the second a confidence interval of [.8016,.8254], meaning that the differences are not statistically different.

## 4.6 Tools and Resources used

The organizers asked participants to submit a description file, special emphasis on the tools and resources that they used. Table 4.3 shows in a simplified way the tools and resources used by those participants that did submit a valid description file. In the last row, the totals show that WordNet was the most used resource, followed by monolingual corpora and Wikipedia. Acronyms, dictionaries, multilingual corpora, stopword lists and tables of paraphrases were also used.

Generic NLP tools like lemmatization and PoS tagging were widely used, and to a lesser extent, parsing, word sense disambiguation, semantic role labeling and time and date resolution (in this order). Knowledge-based and distributional methods got used nearly equally, and to a lesser extent, alignment and/or statistical machine translation software, lexical substitution, string similarity, textual entailment and machine translation evaluation software. Machine learning was widely used to combine and tune components.

<sup>4</sup>We run the random baseline 10 times. The mean is reported here. The standard deviation is 0.0005



Several less used tools were also listed but were used by three or less systems.

The top scoring systems tended to use most of the resources and tools listed (*UKP*, *Takelab*), with some notable exceptions like *Sgjimenez* which was based on string similarity.

For a more detailed analysis, the reader is directed to the papers of the participants in this volume.

## 4.7 Best three systems

This section briefly describes the three best systems of the competition.

### 4.7.1 baer/run2

This system uses a simple log-linear regression model, trained on the training data, to combine multiple text similarity measures of varying complexity. They first run different similarity measures separately. After that, they use the resulting scores as features for a machine learning classifier.

The system is based on DKPro3, a collection of software components for natural language processing built upon the Apache UIMA framework. During the pre-processing phase, they tokenize the input texts and lemmatize using the Tree-Tagger implementation (Schmid, 1994). For some measures, they also apply a stopwords filter.

In the next step, they compute similarity scores, generating score vectors which served as features. Then, they perform a feature combination using the pre-computed similarity scores, and combined their log-transformed values using a linear regression classifier from the WEKA toolkit (Hall et al., 2009). They trained the classifier on the training datasets of the STS task.

After that, they applied a post-processing filter which stripped all characters off the texts which are not in the character range.

Finally, during the development cycle, features which achieved the best performance on the training data were identified.

### 4.7.2 jan snajder/run1

This system uses supervised regression with support vector regression (SVR) as a learning model, exploiting different feature sets and SVR hyperparameters.

Firstly, they perform many preprocessing steps for cleanign and normalizing the data. This process includes tokenization, part-of-speech tagging and stop-word removal.

Secondly, they use many features previously seen in paraphrase classification (Michel et al., 2011). Several features are based on consecutive unigrams, bigrams, and trigrams overlap. In addition to the overlap of consecutive ngrams, they also compute the skip bigram and trigram overlap. To allow for some lexical variation, they use WordNet to assign partial scores to words that are not common to both sentences.

They also give more importance to words bearing more content, by computing the frequency of the words in the corpus. They used the Google Books Ngrams (Michel et al., 2011) to obtain word frequencies because of its excellent word coverage for English. Additionally, they measure the similarity between sentences using the semantic alignment of lemmas, in a similar way of a previous research by (Lavie and Denkowski, 2009), who proposed a similar alignment strategy for machine translation evaluation.

In the next step, they use dependency parsing to identify the lemmas with the corresponding syntactic roles in the two sentences. They also compute the overlap of the dependency relations of the two sentences.

Finally for each of the provided training sets they trained a separate Support Vector Regression (SVR) model using LIBSVM (Chang and Lin, 2011).

### 4.7.3 `sgjimenezv/run1`

This team present an approach for the construction of text similarity functions using a parameterized resemblance coefficient in combination with a softened cardinality function called soft cardinality. Classical cardinality counts the number of elements which are not identical in a set, soft cardinality uses an auxiliary inter-element similarity function to make a soft count. For instance, the soft cardinality of a set with two very similar (but not identical) elements should be a real number closer to 1.0 instead of 2.0.

This approach provides a recursive model, varying levels of granularity from sentences to characters. Therefore, the model was used to compare sentences divided into words, and in turn, words divided into q-grams of characters. They observed that a performance correlation function in a space defined by all parameters was relatively smooth and had a single maximum achievable by “hill climbing.” The system used only surface text information, a stop-word remover, and a stemmer to tackle the semantic text similarity task.

## 4.8 Our system

Although it was not the primary objective of this thesis, a system was also presented to the competition. The system is based on Moses, an Statistical Machine Translation (SMT) system.

The first step was the lemmatization and the part-of-speech tagging of the phrases. After obtaining the lemma and PoS a stop-word-list was used to remove the stop-words.

In the second step we built phrase-tables to store the probabilities for a given word to be translated as another (for example, two words would have probability 1 if they are the same word or synonyms). Different phrase-tables were constructed with different values depending on whether they are the same word, the same part-of-speech or completely different.

In the last step, we introduced to Moses a source sentence and a target sentence. A constraint was applied to Moses, so it was forced to translate the source sentence to the target sentence. That is, the application returns us the possibility of our target translation to be the correct translation:  $P(\text{target}|\text{source})$ . This probability is used as similarity value.

The system was not trained nor adjusted for the train datasets. Despite this, our system slightly improved the baseline system. One goal for the future is to improve the system for future editions of semeval.

## 4.9 Conclusions

This chapter presents the SemEval 2012 pilot evaluation exercise on Semantic Textual Similarity. A simple definition of STS beyond the likert-scale was set up, and a wealth of annotated data was produced. The similarity of pairs of sentences was rated on a 0-5 scale (low to high similarity) by human judges using Amazon Mechanical Turk. The dataset includes 1500 sentence pairs from MSRpar and MSRvid (each), ca. 1500 pairs from WMT, and 750 sentence pairs from a mapping between OntoNotes and WordNet senses. The correlation between non-expert annotators and annotations from the authors is very high, showing the high quality of the dataset. The dataset was split 50% as train and test, with the exception of the surprise test datasets: a subset of WMT from a different domain and the OntoNotes-WordNet mapping. All datasets are publicly available.<sup>5</sup>

The exercise was very successful in participation and results. 35 teams participated, submitting 88 runs. The best results scored a Pearson correlation over 80%, well beyond a simple lexical baseline with 31% of correlation. The metric for evaluation was not completely satisfactory, and three evaluation metrics were finally published. We discuss the shortcomings of those measures.

There are several tasks ahead in order to make STS a mature field. The first is to find a satisfactory evaluation metric. The second is to analyze the definition of the task itself, with a thorough analysis of the definitions in the likert scale.

We would also like to analyze the relation between the STS scores and the paraphrase judgements in MSR, as well as the human evaluations in WMT. Finally, we would also like to set up an open framework where NLP components and similarity algorithms can be combined by the community. All in all, we would like this dataset to be the focus of the community working on algorithmic approaches for semantic processing and inference at large.

---

<sup>5</sup><http://www.cs.york.ac.uk/semeval-2012/task6/>

Run	ALL	Rank	ALLnorm	Rank	Mean	Rank	MSRpar	MSRvid	SMT-eur	On-WN	SMT-news
00-baseline/run1	.3110	87	.6732	85	.4356	70	.4334	.2996	.4542	.5864	.3908
aca08ls/run1	.6485	34	.8238	15	.6100	18	.5166	.8187	.4859	.6676	.4280
aca08ls/run2	.7241	17	.8169	18	.5750	38	.5166	.8187	.4859	.6390	.2089
aca08ls/run3	.6054	48	.7946	44	.5943	27	.5460	.7241	.4858	.6676	.4280
acaputo/run1	.6141	46	.8027	38	.5891	31	.4542	.7673	.5126	.6593	.4636
acaputo/run2	.6221	44	.8079	30	.5728	40	.3886	.7908	.4679	.6826	.4238
acaputo/run3	.6285	41	.7951	43	.5651	45	.4128	.7612	.4531	.6306	.4887
baer/run1	.8117	4	.8559	4	.6708	4	.6821	.8708	.5118	.6649	.4672
baer/run2	<b>.8239</b>	1	<b>.8579</b>	2	<b>.6773</b>	1	.6830	.8739	.5280	.6641	.4937
baer/run3	.7790	8	.8166	19	.4320	71	.6830	.8739	.5280	-.0620	-.0520
croce/run1	.7474	13	.8292	12	.6316	10	.5695	.8217	.5168	.6591	.4713
croce/run2	.7475	12	.8297	11	.6323	9	.5763	.8217	.5102	.6591	.4713
croce/run3	.6289	40	.8150	21	.5939	28	.4686	.8027	.4574	.6591	.4713
csjxu/run1	.6528	31	.7642	59	.5492	51	.4728	.6593	.4835	.6196	.4290
danielcer/run1†	.6354	38	.7212	70	.4848	66	.3795	.5350	.4377	.6052	.4164
danielcer/run2†	.4229	77	.7160	72	.5044	62	.4409	.4698	.4558	.6468	.4769
danielcer/run3†	.5589	55	.7807	55	.4674	67	.4374	.8037	.3533	.3077	.3235
davide_buscaldi/run1	.4280	76	.7379	65	.5009	63	.4295	.6125	.4952	.5387	.3614
davide_buscaldi/run2	.4813	68	.7569	61	.5202	58	.4171	.6728	.5179	.5526	.3693
davide_buscaldi/run3	.4064	81	.7287	69	.4898	65	.4326	.5833	.4856	.5317	.3480
demetrios_glinos/run1	.3454	83	.6990	81	.2772	87	.1684	.6256	.2244	.1648	.0988
demetrios_glinos/run2	.4976	64	.7160	73	.3215	86	.2312	.6595	.1504	.2735	.1426
demetrios_glinos/run3	.4165	79	.7129	75	.3312	85	.1887	.6482	.2769	.2950	.1336
desouza/run1	.5633	54	.7127	76	.3628	82	.2494	.6117	.1495	.4212	.2439
desouza/run2	.6438	35	.8080	29	.5888	32	.5128	.7807	.3796	.6228	.5474
desouza/run3	.6517	32	.8106	25	.6077	20	.5169	.7773	.4419	.6298	.6085
dvilarioayala/run1	.4997	63	.7568	62	.5260	56	.4037	.6532	.4521	.6050	.4537
dvilarioayala/run2	-.0260	89	.5933	89	.1016	89	.1109	.0057	.0348	.1788	.1964
dvilarioayala/run3	.6630	25	.7474	64	.5105	59	.4018	.6378	.4758	.5691	.4057
enrique/run1	.4381	75	.7518	63	.5577	48	.5328	.5788	.4785	.6692	.4465
enrique/run2	.2791	88	.6694	87	.4286	72	.3861	.2570	.4086	.6006	.5305
enrique/run3	.4680	69	.7625	60	.5615	47	.5166	.6303	.4625	.6442	.4753
georgiana_dinu/run1	.4952	65	.7871	50	.5065	60	.4043	.7718	.2686	.5721	.3505
georgiana_dinu/run2	.4548	71	.8258	13	.5662	43	.6310	.8312	.1391	.5966	.3806
jan_snajder/run1	<b>.8133</b>	3	<b>.8635</b>	1	<b>.6753</b>	2	.7343	.8803	.4771	.6797	.3989
jan_snajder/run2	<b>.8138</b>	2	<b>.8569</b>	3	.6601	5	.6985	.8620	.3612	.7049	.4683
janardhan/run1	.3431	84	.6878	84	.3481	83	.1936	.5504	.3755	.2888	.3387
jhasneha/run1	.6622	27	.8048	34	.5654	44	.5480	.7844	.3513	.6040	.3607
jhasneha/run2	.6573	28	.8083	28	.5755	37	.5610	.7857	.3568	.6214	.3732
jhasneha/run3	.6497	33	.8043	36	.5699	41	.5460	.7818	.3547	.5969	.4137
jotacastillo/run1	.5522	57	.7904	47	.5906	29	.5659	.7113	.4739	.6542	.4253
jotacastillo/run2	.6272	42	.8032	37	.5838	34	.5538	.7706	.4480	.6135	.3894
jotacastillo/run3	.6311	39	.7943	45	.5649	46	.5394	.7560	.4181	.5904	.3746
Konstantin_Z/run1	.5636	53	.8052	33	.5759	36	.4797	.7821	.4576	.6488	.3682
M_Rios/run1	.6397	36	.7187	71	.3825	80	.3628	.6426	.3074	.2806	.2082
M_Rios/run2	.5981	49	.6955	82	.3473	84	.3529	.5724	.3066	.2643	.1164
M_Rios/run3	.5361	59	.6287	88	.2567	88	.2995	.2910	.1611	.2571	.2212
mheilman/run1	.7808	7	.8064	32	.6305	11	.6211	.7210	.4722	.7080	.5149
mheilman/run2	.7834	6	.8089	27	.6399	7	.6397	.7200	.4850	.7124	.5312
mheilman/run3	.4477	73	.7291	68	.5253	57	.5049	.5217	.4748	.6169	.4566
nitish_aggarwal/run1*	.5777	52	.8158	20	.5466	52	.3675	.8427	.3534	.6030	.4430
nitish_aggarwal/run2*	.5833	51	.8183	17	.5683	42	.3720	.8330	.4238	.6513	.4499
nitish_aggarwal/run3	.4911	67	.7696	57	.5377	53	.5320	.6874	.4514	.5827	.2818
nmalandrakis/run1	.6228	43	.8100	26	.5979	23	.5984	.7717	.4292	.6480	.3702
nmalandrakis/run2	.5540	56	.7997	41	.5558	50	.5960	.7616	.2628	.6016	.3446
nmalandrakis/run3	.4918	66	.7646	58	.5061	61	.4989	.7092	.4437	.4879	.2441
parthapakray/run1*	.3880	82	.6706	86	.4111	76	.3427	.3549	.4271	.5298	.4034
rada/run1	.7418	14	.8406	7	.6159	14	.5032	.8695	.4797	.6715	.4033
rada/run2	.7677	9	.8389	9	.5947	25	.5693	.8688	.4203	.6491	.2256
rada/run3	.7846	5	.8440	6	.6162	13	.5353	.8750	.4203	.6715	.4033
sbdlrhmn/run1	.6663	23	.7842	53	.5376	54	.5440	.7335	.3830	.5860	.2445
sbdlrhmn/run2	.4169	78	.7104	77	.4986	64	.4617	.4489	.4719	.6353	.4353
sgjimenezv/run1	.7331	15	.8526	5	<b>.6708</b>	3	.6405	.8562	.5152	.7109	.4833
sgjimenezv/run2	.7107	19	.8397	8	.6486	6	.6316	.8237	.4320	.7109	.4833
siva/run1	.5253	60	.7962	42	.6030	21	.5735	.7123	.4781	.6984	.4177
siva/run2	.5490	58	.8047	35	.5943	26	.5020	.7645	.4875	.6677	.4324
siva/run3	.5130	61	.7895	49	.5287	55	.3765	.7761	.4161	.5728	.3964
skamler_/run1*†	.3129	86	.6935	83	.3889	79	.3605	.5187	.2259	.4098	.3465
sokolov/run1	.6392	37	.7344	67	.3940	78	.3948	.6597	.0143	.4157	.2889
sokolov/run2	.6789	22	.7377	66	.4118	75	.4848	.6636	.0934	.3706	.2455
sokolov/run3	.6196	45	.7101	78	.4131	74	.4295	.5724	.2842	.3989	.2575
spirin2/run1	.4592	70	.7800	56	.5782	35	.6523	.6691	.3566	.6117	.4603
spirin2/run2	.7269	16	.8217	16	.6104	17	.5769	.8203	.4667	.5835	.4945
spirin2/run3	.3216	85	.7857	51	.4376	69	.5635	.8056	.0630	.2774	.2409
sranjans/run1	.6529	30	.8018	39	.6249	12	.6124	.7240	.5581	.6703	.4533
sranjans/run2	.6651	24	.8128	22	.6366	8	.6254	.7538	.5328	.6649	.5036
sranjans/run3	.5045	62	.7846	52	.5905	30	.6167	.7061	.5666	.5664	.3968
tiantianzhu7/run1	.4533	72	.7134	74	.4192	73	.4184	.5630	.2083	.4822	.2745
tiantianzhu7/run2	.4157	80	.7099	79	.3960	77	.4260	.5628	.1546	.4552	.1923
tiantianzhu7/run3	.4446	74	.7097	80	.3740	81	.3411	.5946	.1868	.4029	.1823
weiwei/run1*†	.6946	20	.8303	10	.6081	19	.4106	.8351	.5128	.7273	.4383
yeh/run1†	.7513	11	.8017	40	.5997	22	.6084	.7458	.4688	.6315	.3994
yeh/run2†	.7562	10	.8111	24	.5858	33	.6050	.7939	.4294	.5871	.3366
yeh/run3†	.6876	21	.7812	54	.4668	68	.4791	.7901	.2159	.3843	.2801
ygutierrez/run1	.6630	26	.7922	46	.5560	49	.6022	.7709	.4435	.4327	.4264
ygutierrez/run2	.6529	29	.8115	23	.6116	16	.5269	.7756	.4688	.6539	.5470
ygutierrez/run3	.7213	18	.8239	14	.6158	15	.6205	.8104	.4325	.6256	.4340
yrkakde/run1	.5977	50	.7902	48	.5742	39	.5294	.7470	.5531	.5698	.3659
yrkakde/run2	.6067	47	.8078	31	.5955	24	.5757	.7765	.4989	.6257	.3468

Table 4.1: The first row corresponds to the baseline. **ALL** for overall Pearson, **ALLnorm** for Pearson after normalization, and **Mean** for mean of Pearsons. We also show the ranks for each measure. Rightmost columns show Pearson for each individual dataset. Note: \* system submitted past the 120 hour window, † team involving one of the organizers.

Run	ALL	ALL <sub>w</sub>	MSRpar	MSRpar <sub>w</sub>	MSRvid	MSRvid <sub>w</sub>	SMT-eur	SMT-eur <sub>w</sub>	On-WN	On-WN <sub>w</sub>	SMT-news	SMT-news <sub>w</sub>
davide_buscaldi/run1	.4280	<b>.4946</b>	<b>.4295</b>	.4082	.6125	<b>.6593</b>	.4952	<b>.5273</b>	.5387	<b>.5574</b>	.3614	<b>.4674</b>
davide_buscaldi/run2	.4813	<b>.5503</b>	<b>.4171</b>	.4033	.6728	<b>.7048</b>	.5179	<b>.5529</b>	.5526	<b>.5950</b>	.3693	<b>.4648</b>
davide_buscaldi/run3	.4064	<b>.4682</b>	<b>.4326</b>	.4035	.5833	<b>.6253</b>	.4856	<b>.5138</b>	<b>.5317</b>	.5189	.3480	<b>.4482</b>
enrique/run1	<b>.4381</b>	.2615	<b>.5328</b>	.4494	<b>.5788</b>	.4913	<b>.4785</b>	.4660	<b>.6692</b>	.6440	<b>.4465</b>	.3632
enrique/run2	<b>.2791</b>	.2002	<b>.3861</b>	.3802	<b>.2570</b>	.2343	.4086	<b>.4212</b>	<b>.6006</b>	.5947	<b>.5305</b>	.4858
enrique/run3	<b>.4680</b>	.3754	<b>.5166</b>	.5082	<b>.6303</b>	.5588	.4625	<b>.4801</b>	<b>.6442</b>	.5761	<b>.4753</b>	.4143
parthapakray/run1	<b>.3880</b>	.3636	.3427	<b>.3498</b>	<b>.3549</b>	.3353	<b>.4271</b>	.3989	<b>.5298</b>	.4619	<b>.4034</b>	.3228
tiantianzhu7/run1	.4533	<b>.5442</b>	.4184	<b>.4241</b>	<b>.5630</b>	<b>.5630</b>	.2083	<b>.4220</b>	.4822	<b>.5031</b>	.2745	<b>.3536</b>
tiantianzhu7/run2	.4157	<b>.5249</b>	.4260	<b>.4340</b>	.5628	<b>.5758</b>	.1546	<b>.4776</b>	.4552	<b>.4926</b>	.1923	<b>.3362</b>
tiantianzhu7/run3	.4446	<b>.5229</b>	.3411	<b>.3611</b>	<b>.5946</b>	.5899	.1868	<b>.4769</b>	.4029	<b>.4365</b>	.1823	<b>.4014</b>

Table 4.2: Results according to weighted correlation for the systems that provided non-uniform confidence alongside their scores.

	Acronyms	Dictionaries	Distributional thesaurus	Monolingual corpora	Multilingual corpora	Stop words	Tables of paraphrases	Wikipedia	WordNet	Alignment	Distributional similarity	KB Similarity	Lemmatizer	Lexical Substitution	Machine Learning	MT evaluation	MWE	Named Entity recognition	POS tagger	Semantic Role Labeling	SMT	String similarity	Syntax	Textual entailment	Time and date resolution	Word Sense Disambiguation	Other	
aca08ls/run1									x			x	x													x	x	
aca08ls/run2									x			x	x		x												x	x
aca08ls/run3									x			x	x		x												x	x
baer/run1		x	x	x	x			x	x		x	x	x						x									x
baer/run2		x	x	x	x			x	x		x	x	x						x									x
baer/run3		x	x	x	x			x	x		x	x	x						x									x
croce/run1				x							x		x						x									
croce/run2				x							x		x						x									
croce/run3				x							x		x						x									
csjxu/run1								x	x			x							x									
danielcer/run1						x		x					x	x					x									
danielcer/run2						x		x					x	x					x									
danielcer/run3								x	x				x						x									
davide_buscaldi/run1				x					x			x							x									
davide_buscaldi/run2				x					x			x							x									
davide_buscaldi/run3				x					x			x							x									
demetrios_glinos/run1								x			x	x							x	x								x
demetrios_glinos/run2								x			x	x							x	x								x
demetrios_glinos/run3								x			x	x							x	x								x
desouza/run1	x		x			x	x	x			x	x	x	x				x	x									x
desouza/run2			x			x	x	x			x	x	x						x									
desouza/run3			x			x	x	x			x								x									
dvilarinoayala/run1	x												x															
dvilarinoayala/run2								x																				
dvilarinoayala/run3													x															x
jan_snajder/run1	x	x	x		x			x	x		x	x	x	x	x				x									x
jan_snajder/run2				x					x		x	x	x	x					x	x								
janardhan/run1									x				x						x	x								x
jotacastillo/run1	x			x					x				x	x														x
jotacastillo/run2	x			x					x				x	x														x
jotacastillo/run3	x			x					x				x	x														x
Konstantin_Z/run1																												
M_Rios/run1			x								x		x						x	x	x							
M_Rios/run2			x								x		x						x	x	x							
M_Rios/run3			x								x		x						x	x	x							
mheilman/run1				x					x			x	x	x	x													
mheilman/run2				x	x				x			x	x	x	x													x
mheilman/run3				x	x				x			x	x	x														x
parthapakray/run1	x				x				x				x						x	x	x							x
rada/run1									x	x	x	x	x	x														x
rada/run2									x	x	x	x	x	x														x
rada/run3									x	x	x	x	x	x														x
sgjimenezv/run1					x								x	x														
sgjimenezv/run2					x								x	x														
skamler_/run1						x					x		x							x								
sokolov/run1				x							x		x															
sokolov/run2				x							x		x															
sokolov/run3				x							x		x															
spirin2/run1	x			x		x	x												x	x	x	x						x
spirin2/run2	x			x		x	x												x	x	x	x						x
spirin2/run3	x			x		x	x												x	x	x	x						x
srnjans/run1				x					x	x	x	x																x
srnjans/run2				x					x	x	x	x																x
srnjans/run3				x					x	x	x	x																x
tiantianzhu7/run1									x			x																x
tiantianzhu7/run2									x			x																
tiantianzhu7/run3					x				x																			
weiwei/run1	x			x					x			x																
yeh/run1			x						x	x		x	x	x														
yeh/run2			x						x	x		x	x	x														
yeh/run3			x						x	x		x	x	x														
ygutierrez/run1				x					x	x		x	x	x														x
ygutierrez/run2				x					x			x	x															x
ygutierrez/run3				x					x	x		x	x															x
yrkakde/run1									x			x																
Total	8	6	10	33	5	5	9	20	47	7	31	37	49	13	13	4	7	12	43	9	4	13	17	10	5	15	25	

Table 4.3: Resources and tools used by the systems that submitted a description file. Leftmost columns correspond to the resources, and rightmost to tools, in alphabetic order.

## Chapter 5

# Concluding Remarks and Future Directions

This chapter presents the main conclusions of this work. It summarizes its main contributions and it defines possible future research lines.

Measuring the semantic similarity and the different relationships between terms is a very important task in the area of lexical semantics. There are two main branches to try to solve this problem. The first one is based on techniques that use structures resources like WordNet or Wikipedia. The second ones are based on the use of large monolingual or multilingual corpora.

However, most of the proposed techniques are commonly evaluated on manually created datasets, where the weights returned by the systems are compared with scores assigned by humans. These are very few datasets for Semantic Textual Similarity (STS). In this work we have worked with a newly created dataset, which although it is quite small, has been useful enough to evaluate different similarity algorithms.

Moreover, most of these techniques are applied at word level, and very few at sentence or text level. This is because compositionality, which makes the calculation of similarities between phrases very complex and difficult. Part of this work has consisted in advancing a few steps in this direction. Additionally, in the framework of SemEval-2012 we organized a pilot task for the creation and the evaluation of STS systems capable of working with phrases instead of words.

### 5.1 Main contributions

Our work contributes in different ways to the state of the art on improving and creating new techniques and resources to measure the similarity and relatedness between textual items. In particular, the main contributions of our work can be summarized as follows:

1. We provided an in depth study of the state of the art in the area of semantic textual similarity.

2. We conducted and evaluated an empirical study on measuring the semantic similarity between cultural heritage items.
3. Five new datasets have been created. These datasets have been used to evaluate a large set of systems participating in the Semantic Textual Similarity (STS) competition.
4. We also explored different evaluation measures for STS.

The work organizing the Semantic Textual Similarity competition of SemEval also resulted in a publication:

- [Agirre et al., 2012]  
Agirre E., Cer D., Diab M., Gonzalez-Agirre A. (2012). Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*, in conjunction with the First Joint Conference on Lexical and Computational Semantics (\*SEM 2012), Montreal (Canada).

## 5.2 Future work

Despite the good results obtained in the experiments on the Europeana collections, it was not possible to calculate properly the semantic similarity between some pairs of items. The personalized PageRank algorithm works over a graph generated using the WordNet dictionary. Thus, some textual descriptions having proper names, works of art or locations were miss-represented. Moreover, when an item is composed only with named entities not stored in the knowledge base (this occurs especially when we are using only the title or the subject) the algorithm is not able to generate any similarity value.

Therefore, a near future work will consist in enriching our knowledge base with entities extracted from YAGO2 [Johannes Hoffart and Weikum, 2010]<sup>1</sup>. Moreover, now the gold-standard has been enlarged to 400 pairs. We also plan to test our algorithms on this new dataset.

Finally, the STS competition of SemEval-2012 was proposed as a pilot task. This task will be organized again in the coming years<sup>2</sup>, and for the following editions we also expect to find a satisfactory evaluation metric. For future editions, we also expect to improve the datasets and the systems, advancing the state of the art in this research field.

---

<sup>1</sup><http://www.mpi-inf.mpg.de/yago-naga/yago/>

<sup>2</sup>SemEval 2013 plans to organize an a new STS evaluation campaign <http://www.cs.york.ac.uk/semeval-2013/index.php?id=tasks>



# Bibliography

- [Agirre et al., 2012] Agirre, E., Cer, D., Diab, M., and Gonzalez-Agirre, A. (2012). SemEval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012), in conjunction with the First Joint Conference on Lexical and Computational Semantics, SemEval '12*, Montreal, Canada.
- [Agirre et al., 2010] Agirre, E., Cuadros, M., Rigau, G., and Soroa, A. (2010). Exploring Knowledge Bases for Similarity. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC10). European Language Resources Association (ELRA). ISBN: 2-9517408-6-7. Pages 373–377.*”.
- [Agirre and Soroa, 2009a] Agirre, E. and Soroa, A. (2009a). Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, EACL '09*, pages 33–41, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Agirre and Soroa, 2009b] Agirre, E. and Soroa, A. (2009b). Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009)*, Athens, Greece.
- [Agirre et al., 2009] Agirre, E., Soroa, A., Alfonseca, E., Hall, K., Kravalova, J., and Pasca, M. (2009). A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In *Proceedings of annual meeting of the North American Chapter of the Association of Computational Linguistics*.
- [Aletras, 2011] Aletras, N. (2011). *Computing Similarity between Items in a Digital Library of Cultural Heritage*. PhD thesis, Department of Computer Science, The University of Sheffield.
- [Alvarez and Lim, 2007] Alvarez, M. A. and Lim, S. (2007). A Graph Modeling of Semantic Similarity between Words. In *Proceedings of the International Conference on Semantic Computing, ICSC '07*, pages 355–362, Washington, DC, USA. IEEE Computer Society.

- [Baeza-Yates and Ribeiro-Neto, 1999] Baeza-Yates, R. A. and Ribeiro-Neto, B. A. (1999). *Modern Information Retrieval*. ACM Press / Addison-Wesley.
- [Banchs and Costa-jussà, 2011] Banchs, R. E. and Costa-jussà, M. R. (2011). A semantic feature for statistical machine translation. In *Proceedings of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation, SSST-5*, pages 126–134, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Bollegala et al., 2009] Bollegala, D., Matsuo, Y., and Ishizuka, M. (2009). Measuring the similarity between implicit semantic relations using web search engines. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining, WSDM '09*, pages 104–113, New York, NY, USA. ACM.
- [Briscoe and Boguraev, 1989] Briscoe, T. and Boguraev, B. (1989). *Computational lexicography for natural language processing*. Longman Publishing Group, White Plains, NY, USA.
- [Callison-Burch et al., 2007] Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2007). (Meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation, StatMT '07*, pages 136–158.
- [Callison-Burch et al., 2008] Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2008). Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation, StatMT '08*, pages 70–106.
- [Castillo and Cardenas, 2010] Castillo, J. J. and Cardenas, M. E. (2010). Using sentence semantic similarity based on WordNet in recognizing textual entailment. In *Proceedings of the 12th Ibero-American conference on Advances in artificial intelligence, IBERAMIA'10*, pages 366–375, Berlin, Heidelberg. Springer-Verlag.
- [Chen and Dolan, 2011] Chen, D. L. and Dolan, W. B. (2011). Collecting Highly Parallel Data for Paraphrase Evaluation. In *Proceedings of the 49th Annual Meetings of the Association for Computational Linguistics (ACL)*.
- [Chen et al., 2006] Chen, H.-H., Lin, M.-S., and Wei, Y.-C. (2006). Novel association measures using web search with double checking. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, ACL-44*, pages 1009–1016, Stroudsburg, PA, USA. Association for Computational Linguistics.

- [Church and Hanks, 1990] Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16:22–29.
- [Cuadros and Rigau, 2008] Cuadros, M. and Rigau, G. (2008). KnowNet: Building a Large Net of Knowledge from the Web. In Scott, D. and Uszkoreit, H., editors, *COLING*, pages 161–168.
- [Deerwester et al., 1990] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, 41(6):391–407.
- [Dennis, 1964] Dennis, S. (1964). The construction of a thesaurus automatically from a sample of text. *Statistical association methods for mechanized documentation, symposium proceedings (Miscellaneous publication 269)*. Washington, DC: National Bureau of Standards.
- [Dolan et al., 2004] Dolan, B., Quirk, C., and Brockett, C. (2004). Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, page 350.
- [Dunning, 1993] Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.*, 19:61–74.
- [Fellbaum, 1998a] Fellbaum, C. (1998a). *WordNet. An Electronic Lexical Database*. Language, Speech, and Communication. The MIT Press.
- [Fellbaum, 1998b] Fellbaum, C. (1998b). *WordNet: An Electronic Lexical Database*. MIT Press.
- [Fern and Stevenson, 2009] Fern, S. and Stevenson, M. (2009). A Semantic Similarity Approach to Paraphrase Detection.
- [Finkelstein et al., 2002] Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2002). Placing Search in Context: The Concept Revisited. *ACM Transactions on Information Systems*, 20(1):116–131.
- [Gabrilovich and Markovitch, 2007] Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on Artificial intelligence, IJCAI'07*, pages 1606–1611, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- [Grefenstette, 1994] Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Norwell, MA, USA.
- [Haveliwala, 2002] Haveliwala, T. H. (2002). Topic-sensitive PageRank. In *WWW '02*, pages 517–526, New York, NY, USA. ACM.

- [Hirst and St-Onge, 1998] Hirst, G. and St-Onge, D. (1998). *WordNet: An Electronic Lexical Database - Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms*, in *Wordnet: An Electronic Lexical Database*, chapter 13, pages 305–332. MIT Press.
- [Hliaoutakis et al., 2006] Hliaoutakis, A., Varelas, G., Voutsakis, E., Petrakis, E. G. M., and Milios, E. (2006). Information Retrieval by Semantic Similarity. In *Intern. Journal on Semantic Web and Information Systems (IJSWIS)*, 3(3):55–73, July/Sept. 2006. *Special Issue of Multimedia Semantics*.
- [Hovy et al., 2006] Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). OntoNotes: The 90% Solution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*.
- [Hughes and Ramage, 2007] Hughes, T. and Ramage, D. (2007). Lexical semantic relatedness with random graph walks. In *In Proceedings of EMNLP-CoNLL*, pages 581–589.
- [Jiang and Conrath, 1997] Jiang, J. J. and Conrath, D. W. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy.
- [Johannes Hoffart and Weikum, 2010] Johannes Hoffart, Fabian Suchanek, K. B. and Weikum, G. (2010). YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipediag. Technical report, Research Report MPI-I-2010-5-007, Max-Planck-Institut für Informatik, November.
- [Jones, 1972] Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21.
- [Kageura and Umino, 1996] Kageura, K. and Umino, B. (1996). Methods of automatic term recognition: a review. *Terminology*, 3(2):259–289.
- [Landauer and Dumais, 1997] Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato’s problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. In *Psychological Review*, 104(2), pages 211–240.
- [Leacock and Chodorow, 1998] Leacock, C. and Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. In Fellbaum, C., editor, *MIT Press*, pages 265–283, Cambridge, Massachusetts.
- [Lee et al., 2005] 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*, pages 1254–1259, Mahwah, NJ.

- [Lesk, 1986] Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from a ice cream cone. In *Proceedings of SIGDOC'86*.
- [Li et al., 1995] Li, X., Szpakowicz, S., and Matwin, S. (1995). A WordNet-based algorithm for word sense disambiguation. In *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2, IJCAI'95*, pages 1368–1374, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- [Li et al., 2006] 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 Transactions on Knowledge and Data Engineering*, 18(8):1138–1150.
- [Lin, 1997] Lin, D. (1997). Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, ACL '98*, pages 64–71, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Luhn, 1957] Luhn, H. P. (1957). A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development*, 1(4):309–317.
- [Makoto et al., 1976] Makoto, N., Mikio, M., and Hiroyuki, I. (1976). An Automatic Method of the Extraction of Important Words from Japanese Scientific Documents. *Information processing in Japan*, 16:83–88.
- [Manning and Schütze, 1998] Manning, C. D. and Schütze, H. (1998). *Foundations of Statistical Natural Language Processing*. MIT Press.
- [Miller et al., 1991] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K., and Teng, R. (1991). Five Papers on WordNet. *Special Issue of the International Journal of Lexicography*, 3(4):235–312.
- [Mohler et al., 2011] Mohler, M., Bunescu, R., and Mihalcea, R. (2011). Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT '11*, pages 752–762, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Nagwani and Verma, 2011] Nagwani, N. K. and Verma, S. (2011). A Frequent Term and Semantic Similarity based Single Document Text Summarization Algorithm. *International Journal of Computer Applications*, 17(2):36–40. Published by Foundation of Computer Science.

- [Page et al., 1999] Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank Citation Ranking: Bringing Order to the Web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.
- [Pantel and Lin, 2000] Pantel, P. and Lin, D. (2000). An unsupervised approach to prepositional phrase attachment using contextually similar words. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, ACL '00*, pages 101–108, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Patwardhan et al., 2003] Patwardhan, S., Banerjee, S., and Pedersen, T. (2003). Using Measures of Semantic Relatedness for Word Sense Disambiguation. In Gelbukh, A. F., editor, *CICLing*, volume 2588 of *Lecture Notes in Computer Science*, pages 241–257. Springer.
- [Pedersen and Patwardhan, 2004] Pedersen, T. and Patwardhan, S. (2004). Wordnet::similarity - measuring the relatedness of concepts. pages 1024–1025.
- [Rapp, 1999] Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 11–14, Maryland.
- [Rapp, 2004] Rapp, R. (2004). A freely available automatically generated thesaurus of related words. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pages 395–398, Lisboa, Portugal.
- [Resnik, 1995] Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence, IJCAI*.
- [Rigau et al., 1998] Rigau, G., Rodríguez, H., and Agirre, E. (1998). Building Accurate Semantic Taxonomies from Monolingual MRDs. In *Proceedings of COLING/ACL*, Montréal, Canada.
- [Robertson and Jones, 1976] Robertson, S. E. and Jones, K. S. (1976). Relevance weighting of search terms. *J. Am. Soc. Inf. Sci.*, 27(3):129–146.
- [Rubenstein and Goodenough, 1965] Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. *Commun. ACM*, 8(10):627–633.
- [Sahami and Heilman, 2006] Sahami, M. and Heilman, T. D. (2006). A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th international conference on World Wide Web, WWW '06*, pages 377–386, New York, NY, USA. ACM.

- [Salton and Buckley, 1988] Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. In *Proceedings of Information Processing and Management*, pages 513–523.
- [Salton and McGill, 1986] Salton, G. and McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA.
- [Salton et al., 1975] Salton, G. M., Wong, A. K. C., and Yang, C.-S. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11):613–620.
- [Schütze, 1998] Schütze, H. (1998). Automatic word sense discrimination. *Comput. Linguist.*, 24:97–123.
- [Schütze, 1995] Schütze, H. (1995). Distributional Part-of-speech Tagging. In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics, EACL*, Dublin, Ireland.
- [Smadja, 1993] Smadja, F. (1993). Retrieving collocations from text: Xtract. *Comput. Linguist.*, 19:143–177.
- [Strube and Ponzetto, 2006] Strube, M. and Ponzetto, S. (2006). WikiRelate! Computing semantic relatedness using Wikipedia. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1419. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.
- [Turney, 2006] Turney, P. D. (2006). Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.
- [Ukkonen, 1985] Ukkonen, E. (1985). Algorithms for Approximate String Matching. *Information and Control*, 64:110–118.
- [Versley, 2008] Versley, Y. (2008). Decorrelation and shallow semantic patterns for distributional clustering of nouns and verbs. In *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics*, pages 55–62, Hamburg, Germany.
- [Wilks et al., 1996] Wilks, Y., Sinator, B., and Guthrie, L. (1996). *The Grammar of Sense: Is Word-sense Tagging Much More than Part-of-speech Tagging*. MIT Press.
- [Yang and Powers, 2005] Yang, D. and Powers, D. M. W. (2005). Measuring semantic similarity in the taxonomy of WordNet. In *Proceedings of the Twenty-eighth Australasian conference on Computer Science - Volume 38, ACSC '05*, pages 315–322, Darlinghurst, Australia, Australia. Australian Computer Society, Inc.
- [Yeh et al., 2009] Yeh, E., Ramage, D., Manning, C. D., Agirre, E., and Soroa, A. (2009). WikiWalk: random walks on Wikipedia for semantic relatedness. In *Proceedings of the 2009 Workshop on Graph-based*

*Methods for Natural Language Processing*, TextGraphs-4, pages 41–49, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Zernik, 1991] Zernik, U. (1991). *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*. Lawrence Erlbaum Associates.