# Expert and Corpus-Based Evaluation of a 3-Space Model of Conceptual Blending

**Donny Hurley** and **Yalemisew Abgaz** and **Hager Ali** and **Diarmuid O'Donoghue**[1]

**Abstract.** This paper presents the 3-space model of conceptual blending that estimates the figurative similarity between Input spaces 1 and 2 using both their analogical similarity and the inter-connecting Generic Space. We describe how our Dr Inventor model is being evaluated as a model of lexically based figurative similarity. We describe distinct but related evaluation tasks focused on 1) identifying novel and quality analogies between computer graphics publications 2) evaluation of machine generated translations of text documents 3) evaluation of documents in a plagiarism corpus. Our results show that Dr Inventor is capable of generating novel comparisons between publications but also appears to be a useful tool for evaluating machine translation systems and for detecting and assessing the level of plagiarism between documents. We also outline another more recent evaluation, using a corpus of patent applications.

## 1 Introduction

The Dr Inventor [**1**] system was developed with the specific objective of identifying creative [**2**] analogies between publications from the discipline of computer graphics. The primary focus of Dr Inventor is to identify similarities between graphics publications such that, when these are presented to computer graphics experts will (frequently) cause creative insight in the user, by highlighting some un-noticed similarities. Dr Inventor is focused on identifying analogies between a user's publication and other papers that typically arise from a different topic (and year) within computer graphics. Early results show that the similarities identified by Dr Inventor will almost always suggest novel and identified source papers that generally would not be read by the user.

As well as being a tool to inspire its users' creativity, Dr Inventor aims to assess the novelty of a submitted document in relation to the other documents contained within its corpus. For example, its users may wish to assess the novelty of an Abstract before writing the full paper. Alternatively, a novice author may write the Abstract of a paper and then use Dr Inventor to identify a similar publication from a different topic, using this paper as a guide to writing their own full paper.

This paper assesses Dr Inventor on challenges related to identifying highly similar or quite similar documents. For example, we wish to assess its ability to quantify the similarity between different versions of the same document. Our focus in this paper is on the metrics used by Dr Inventor and how well they quantify the similarity between highly similar documents and even different versions of the same document. So this paper represents an evaluation of the system at a task that differs from its primary objective. However, the first result we shall discuss relies on human expertise of senior researchers to perform the evaluation.

The paper begins with a brief overview of approaches to retrieving similar texts. We then describe a model of analogy-based similarity before describing the Dr Inventor model for discovering novel and useful analogies between computer graphics publications. Our evaluation and results are then presented in three parts: 1) expert evaluation of the two creative analogies discovered from a corpus of papers from the SIGGRAPH[2] conference series. 2) evaluation of machine generated forward-backward translations 3) evaluation of results for a plagiarism corpus. The paper finishes with some general remarks and conclusion on the evaluation of Dr Inventor.

## 2 Document Comparison

Identifying similarities between text-based documents has long been the subject of interest to artificial intelligence. Many approaches have been explored, with some of the more popular approaches being TF-IDF [**3**], LSA [**4**] and many others with many of these approaches being based on word distribution based document representations. Some of the inherent problems with such approaches are discussed in [**5**].

An alternative approach to graph-based document similarity is described in [**5**]. Our approach differs from this in a number of specific regards. Firstly, Dr Inventor's graphs are derived from the output produced by this GATE parser, whereas [**5**] does not use a parser. We do not use external resources to expand the information contained within a document, using the documents as they are presented to perform the similarity assessment. Dr Inventor is based on a cognitive model of figurative thinking, aimed at identifying similarities that are arguably even more abstract and those identified by [**5**]. Our approach looks for figurative similarities that are variously referred to as metaphors, analogies or conceptual blends.

## 3 Analogy and Conceptual Blending

The approach explored and evaluated in this paper is derived from a cognitive model of people's ability to think figuratively, using two distinct systems of information. At its heart lies the computational model of Gentner's [**6**] influential Structure Mapping Theory (SMT), which posits that many figurative comparisons are best understood by identifying the largest common sub-graph between two systems of information. SMT is a 2-space model that explains why two semantically different concepts can be placed in correspondence

[1] Computer Science department, Maynooth University, Ireland, email: donny.hurley@nuim.ie

[2] http://www.siggraph.org/

between two documents, in SMT it is the topology of information that becomes the prime driver in determining the degree of similarity between two documents - this point shall be highlighted later.

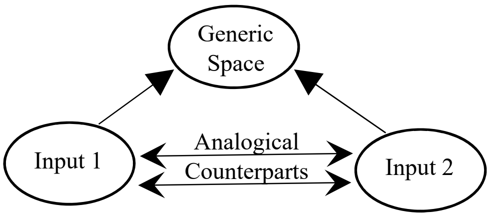


**Figure 1.** A 3-Space model of Blending

Dr Inventor incorporates SMT into a partial implementation of conceptual blending (or Conceptual Integration Networks) [**7**,**8**]. We use conceptual blending theory to extend our 2-space model of analogy, introducing the generic space that represents the abstract commonality between two mapped (and potentially semantically different) concepts or relations. The implementation of SMT used to identify the counterpart projection between the inputs, basing the counterparts on the analogical similarity between them (forming what we call analogical counterparts)..

A figurative comparison that is common in some cultures compares using your legs (for walking) to a bus (specifically, the “number 11 bus”). This comparison might cause “leg” to be mapped to the concept “bus”.  The lexical database WordNet [**9**] might identify an abstract connecting concept, identifying both instances of "instrumentality".  This abstract concept may then be stored in the generic space and may additionally contribute to evaluating the degree of similarity between the mapped concepts.

Figure 1 outlines the structure of the information used by the central graph-based comparison process of the Dr Inventor model. Dr Inventor compares two text based documents not in terms of their raw textual contents, but instead uses a structured representation derived from a dependency parse of that lexical data – using a parser that has been specifically tailored to the needs of this project.

Firstly, Input 1 and Input 2 are rich, highly structured and complex representations of the contents of the two input documents. Their generation will be outlined in section 4 of this paper. Secondly, the analogical counterparts are identified using an implementation of Structure Mapping Theory [**6**] as supported by the VF2 model of graph matching [**10**].  The output of this mapping phase is a list of paired items between Inputs 1 and 2, based primarily on the *structure* of their representations. Finally, the set of paired items (which may not necessarily be the most semantically similar items) are evaluated by identifying the Generic Space that connect each pair of items, using the WordNet lexical database and the *Lin* [**11**] metric. So the evaluations presented in this paper do not look at the Blended Space, but merely assess the level of similarity that already exists between Inputs one and two.

# 4 Dr Inventor

In this section we describe how information is processed through the Dr Inventor [**1**] system and the results that are found.

## 4.1 Input Data

The Dr Inventor system has as its input a Research Object (RO) [**12**] which, for our purposes, are text based documents. The system is focused on the domain of computer graphics and primarily processes academic papers in this domain. However, an RO can be different types of documents such as psychology material, patents or any other form of text based information. In this paper we will describe the processing that occurs with an academic document and this processing can be performed on any text based documents.

## 4.2 Generating the Research Object Skelton (ROS) Graphs

The Dr Inventor Analogy Blended Creativity (DRI-ABC) model does not work on the RO directly, so for the analogy part of the overall Dr Inventor system we first must process a document and create a Research Object Skelton (ROS). A ROS is an attributed relational graph that contains the core information from a document. At the core of a ROS is the Noun-Verb-Noun type of relations (or Concept-Relation-Concept) and this enables the application of Structure Mapping Theory [**6**] of analogy formation. This requires the extraction of the text based information and so the first step required in processing the document is Text Mining.

### *4.2.1 Text Mining Framework*

A RO is typically in the form of a paper in PDF or text format. To generate a ROS it is necessary to extract the different words, find the dependency relations between the words and attach part of speech tags to each word. Additionally, PDF documents introduce further problems in simply extracting sentences; problems arising from the layout, text flow, images, and equations contained within the PDF.

The extraction of subject-verb-object triples from the textual contents of papers is supported by the Dr Inventor Framework [**13**]. This pipeline of scientific text mining modules is distributed as a stand-alone Java library[3] that exposes an API useful to trigger the analysis of articles as well as to easily retrieve the results. For PDF papers the pipeline invokes the PDFX online Web service[4] [**14**] where the paper is converted into an XML document. Core elements such as the title, authors, abstract and the bibliographic entries are identified.

The Noun-Verb-Noun structure is found within individual sentences and sentences are identified by a Sentence Splitter specifically customised to the idiosyncrasies of scientific discourse. For each sentence, a dependency tree is built using a customised version of [**15**], a Citation-aware dependency parser.  The dependency tree identifies types of words (Noun, Verb, Adjective etc.) as well as the types of relationships (subject, object, modifier of nominal etc.). These are used to build the Noun-Verb-Noun structure for the ROS. Additionally the framework identifies co-referent chains in the document, identifying co-referencing words possibly across sentences. This address issues with words such as *it, he, she etc.*

Another feature of the framework is a trainable logistic regression Rhetorical classifier was developed which assigns to each sentence of a paper a rhetorical category (i.e. Background, Approach, Challenge, Outcome and Future Work) used in gold standard

[3] The Dr Inventor Text Mining Framework Java library can be downloaded at: http://backingdata.org/dri/library/

[4] http://pdfx.cs.man.ac.uk/

manually annotated Dr Inventor Corpus [**16**]. Rather than attempting to find analogies between full papers, the rhetorical categories may be used to find analogies in smaller sections of papers, for example is there an analogy between the *background* of one paper and the *background* of another.

#### *4.2.2 ROS Generation from Text Mining Framework Results*

The ROS is constructed by considering the dependency tree formed for each sentence in the publication. As in Agarwal et al. [**17**] a set of rules is applied to these trees, generating connected triples of nouns and verbs. One of the key properties of the ROS graphs is that multiple mentions of the same concept are uniquely represented. This is done either from the co-reference resolution of the text mining framework or by simply joining nodes that have the same word. Relation nodes, i.e. the verbs, can appear multiple times in the ROS.

Each node has an *attribute* of "type" (i.e. noun, verb) and nodes are "tagged" with the rhetorical categories as discussed in the previous section. The format of the ROS was chosen to allow relationships between relations, i.e. second-order or causal relationships between nodes. In the future, when causal relationships are identified by the Text Mining Framework, these nodes will be included in the ROS. The graph database Neo4j[5] uses attributed relational graphs as its representation and as such Dr Inventor uses it for storage of the ROS.

### 4.3 Finding Analogous Document

After storing a collection of ROs in the form of ROS graphs, we want to find the most analogous paper given a chosen target paper. We achieve this by finding (and rating) mappings for the target paper with every other paper contained in the database and choosing the mapping with the highest score. We will now discuss briefly how a mapping is found between one pair of papers.

#### *4.3.1 ROS Mapping*

The generated mapping adheres to Gentner's structure mapping theory [**6**] and its systematicity principle and 1-to-1 mapping constraint. Mapping rules and constraints discussed in [**18**] incorporating both structural mapping and semantic aspects are also utilised. We say that a source graph and a target graph are mapped.

Firstly, the structural mapping between two ROS graphs is based on: 1) graph structure, 2) conceptual structure. Graph structure focuses on identifying isomorphic graphs. Specifically, find the largest isomorphic subgraph of the target in the source. Conceptual structure addresses the conceptual similarity between the nodes and edges that are to be paired by the mapping process [**19**,**20**]. A customised version of the graph matching algorithm VF2 [**10**] is used along with three chosen constraints on the mapping.

Secondly, semantic similarity is used during the computation and the selection phase of candidate pairs. Whenever we encounter two or more candidate pairs that satisfy the structural constraints, rather than selecting a random pair, we select a pair using its semantic similarity. Semantic similarity is calculated using dictionary-based approach. Our semantic mapping utilizes the Lin similarity measure [**11**], which is based on WordNet [**9**], to calculate the similarity between source nodes and target nodes of similar type.

#### *4.3.2 Mapping Metrics*

To select the most analogous source paper for a given target paper we must have some way to rate the mappings. We use a unified metric that combines a structural similarity score with a semantic similarity score to have an overall Unified Analogy Similarity (AS).

Jaccard's coefficient [**21**] is used to measure the structural similarity. The coefficient is used to measure the similarity between two finite sets. The mapping between two graphs is effectively the intersection between the two sets of nodes for the source and the target. As such, the Jaccard's coefficient can be applied. The Jaccard's coefficient has a value between 0 and 1, where if it has value 1 the two ROS graphs are identical and if it is 0 then there is no mapping between the two ROS graphs. Jaccard's coefficient gives an estimate of *how much* of the graphs have been mapped.

For the semantic similarity score we use the same Lin metric as used in the semantic mapping. The *Lin* metric always gives a value between 0 and 1. We calculate the overall semantic similarity of the mapping by getting the average semantic similarity of all paired items in the mapping.

The Unified Analogy Similarity score is calculated by multiplying the Jaccard's coefficient by the Semantic Similarity score giving a value between 0 and 1. After finding the scores for mappings of all source papers with a given target paper, we select the most analogous source paper by whichever has the highest unified analogy similarity.


**Figure 2**. Dr Inventor Paper Processing System

### 4.4 Additional Processing

The above has described the analogy component of the Dr Inventor system. Further processing is done on Computer Graphics papers as part of the overall system. Information is extracted such as topic lists, key words, links between citations, visualisation of similarity between documents and more, as well as a user interface is done by the system, however, this is outside the scope of this paper which is focused on the analogy part of the process.

## 5 Finding and Evaluating a Computer Graphics Analogy

To test and evaluate the DRI-ABC system we created a corpus of 957 papers from the SIGGRAPH computer graphics conference. This is one of the top ranked computer graphics conferences in the world. These papers went through the Text-Mining Framework and the ROS generation components of the Dr Inventor system. Papers were included from the years 2002 to 2011 and included many different sub-topics within this discipline. A small and random selection of papers were chosen to serve as target papers and we found analogous source papers. In this section, we will discuss two

[5] http://www.neo4j.com

of the analogies found. The mapping was performed between the (lexical) *abstract* of each paper combined with the rhetorical category of *background* for each paper and performing the mapping only between these sections of each paper.

In generating the two analogies discussed below, all other papers in the corpus were mapped with the target. From the resulting 956 analogies, the analogy metrics were used to choose only the best source analog for the presented target. Early testing showed that there is frequently an exponential distribution in the quality of the analogical comparisons we discovered (as quantified by the analogy metrics). For this and other reasons, we do not expect Dr Inventor to always find creative analogies for a presented paper. So, in this paper we discuss the two best analogies discovered from a list of the 10 best analogies discovered.

We will first briefly discuss the Target Paper and what the paper is about. Then we will briefly discuss the Source Paper that was chosen by the system. Finally we will talk about feedback from the analogy. This will be qualitative feedback from a senior professor in computer graphics and then quantitative ratings from multiple computer graphics researchers. Each evaluator spent around 50 minutes evaluating each analogy and they were rated on three properties, on a scale from 1-5, 1) novelty 2) usefulness and 3) challenging the normal view of the topic.

## 5.1 First Analogy

### *5.1.1 Target Paper*

The target paper we will discuss is "*Linear Combination of Transformations*" by Marc Alexa which appeared in SIGGRAPH 2002. A brief description of the paper is: This paper's problem is trying to transform a 3D model. The problem is that transforming a 3D model is based on matrix or quaternion operations and these operations are not commutative. The proposed solution is to break each transformation matrix into smaller parts and perform them alternatively and thus the linear combination of smaller matrix transformations is closer to being commutative. Figure 2 shows the topics the paper is contained within (Interaction). This image is generated by Dr Inventor.

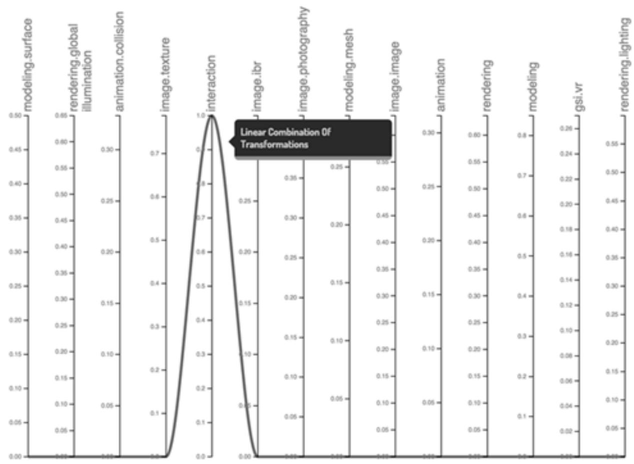


**Figure 3.** Target Paper 1 Topics

### *5.1.2 Source Paper*

Searching through the full corpus of 957 papers, the paper chosen with the highest Analogy Similarity score was "*Gaussian KD-Trees for Fast High-Dimensional Filtering*" by Andrew Adams et al which appeared in SIGGRAPH 2009. A brief description of the paper is: The paper presents an algorithm to accelerate a broad class of non-linear filters. The problem is non-linear filters scale poorly with filter size. The proposed solution it to propose a new Gaussian *kd*-tree, which sparsely represents the high-dimensional space as values stored at points. Figure 3 shows that the paper is contained in the topics Image, Photography and Animation.Collision.

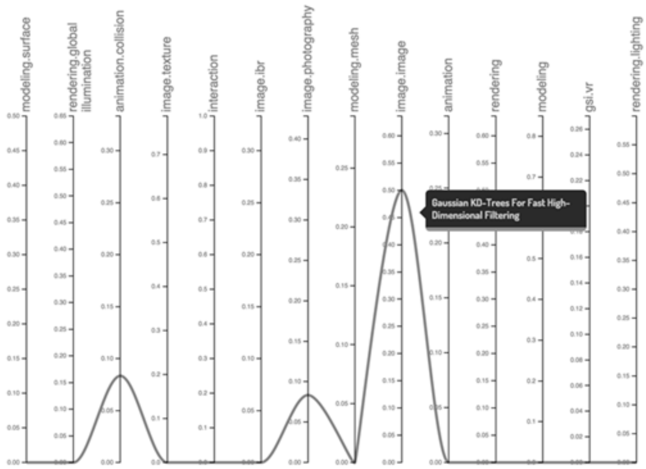


**Figure 4.** Source Paper 1 Topics

### *5.1.3 Analogy Feedback*

A senior professor in Computer Graphics examined these two papers after the system identified them. He considered the two papers to be very analogous and promising. As part of the mapping, the term "matrices" in the target paper was mapped to the term "filter" in the source paper. This suggested that the manipulations applied to matrices can be applied to filters and vice-versa. To show how Dr Inventor could be applied as a Creativity Support Tool this suggested new research ideas that could be further explored. Such as, can we break down image filters into small parts and perform them alternately as was done to the matrices in the analogous paper. Or cascade image filtering and their commutativity.

Two of the interesting things about the found analogy are the differences in the year (2002 and 2009) and also the topics each paper is contained in. They are somewhat dissimilar. This suggests the papers would not usually be compared to one another and they would not typically be papers read when trying to find analogous problems. Dr Inventor is identifying structures not normally considering when trying to find similar papers. Furthermore the conceptual similarity (the semantic similarity between mapped nouns) is 0.37 showing a marked difference between the concepts while a high relational similarity (0.79) was found.

Additionally evaluation of the analogy was performed by 13 evaluators, mostly post-graduate students in computer graphics but also post-doctoral researchers and two senior professors. The average ratings obtained were 4.5 for novelty, 3.7 for usefulness and 4.1 for challenging the normal view of the topic.

## 5.2 Second Analogy

### *5.2.1 Target Paper*

The second target paper is "*Fast Bilateral Filtering for the Display of High-Dynamic-Range Images*" by F Durand and J Dorsey from SIGGRAPH 2002. This paper presents a technique for the display of high-dynamic-range images, which reduces the contrast while preserving details and how poor management of light – under- or over-exposed areas, light behind the main character, etc. – is the single most-commonly-cited reason for rejecting photographs. It has the topics Image Processing and Photograph.

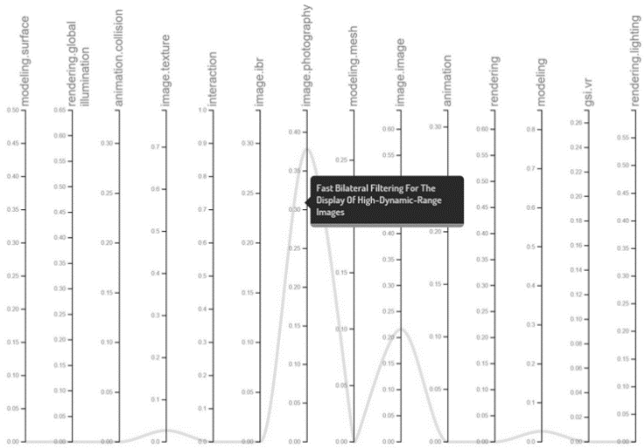


**Figure 5.** Target Paper 2 Topics

### *5.2.2 Source Paper*

The paper with the highest Analogy Similarity score was "*Curve Skeleton Extraction from Incomplete Point Cloud"* by A Tagliasacchi, H Zhang and D Cohen-Or from SIGGRAPH 2009. This paper presents an algorithm for curve skeleton extraction from imperfect point clouds where large portions of the data may be missing. The problem arises from incomplete data during 3D laser scan. The point cloud data contains large holes. The paper has the topics Modeling and Point Cloud.

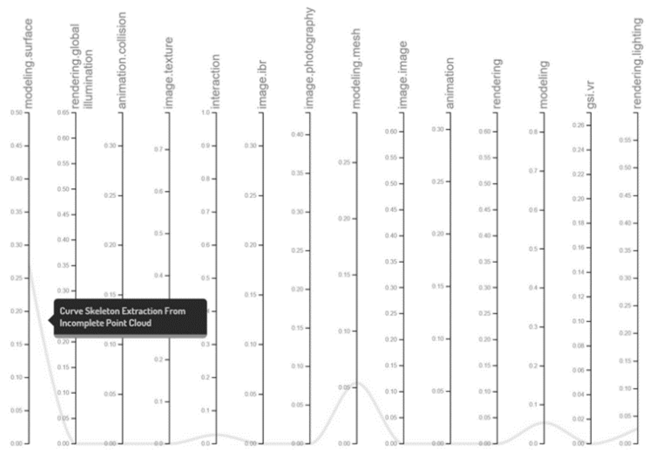


**Figure 6.** Source Paper 2 Topics

### *5.2.3 Analogy Feedback*

A different senior professor provided the qualitative feedback for this analogy. Each paper, when broken down to its basics, is discussing about "missing data" in the image. In the case of the target paper, data about the image is obscured by the contrast of a digital photograph as it cannot as accurately capture the image as the human eye. In the source paper, data points of the 3D image are blocked from being scanned by the lasers. Mappings are found between the term "Hole" in the target paper with "Area" in the source paper. That is "the photo will contain under- and over-exposed areas" is mapped to "data contain large holes caused during 3D laser scan", so Dr Inventor can suggest the similarities between the two paper problems.

The results of this analogy suggested to the professor several possible new ideas for reconstruction of hidden information. How would similar techniques apply to motion capture, missing video data and more.

As in the first example the two papers are found many years apart and the topics they are contained within are not similar. Again, Dr Inventor is finding far analogies that typically would not be found by a normal literature review when attempting to write a research paper. The conceptual similarity was again low (0.37) while the relational similarity was high (0.8).

For the evaluation performed by more researchers, the average ratings were obtained for the same three categories. 4.1 for novelty, 3 for usefulness and 3.3 for challenging norms.

## 6 Further Usage of System

We have described the usage of PDF academic papers through the Dr Inventor system and two of the results found. Additionally, Dr Inventor can be expanded outside its original focus on the domain of computer graphics. ROS graphs can be formed from any text based documents and commonly used plain text files can be processed through the system. We now discuss some of these specific formats that can be used.

We describe the evaluation of Dr Inventor on two tasks that lie beyond the initial scope of this project. Firstly we assess Dr Inventor and particularly its similarity metrics at the task of automatically evaluating the faithfulness of machine translation services. Secondly, we assess it at the task of detecting the degree of similarity between a document and plagiarised versions of those documents. In this section we focus our evaluation on aggregations of results rather than presenting individual comparisons.

## 6.1 Machine Translation Evaluation

Another means of evaluating the DRI-ABC system is to evaluate translations generated by machine translation services. So, this section represents a joint evaluation of DRI-ABC as well as the machine translations themselves. This is searching for a near analogy i.e. generating similar but slightly different versions of the document.

By taking an original document (in English), translating it to the chosen language and then translating this back (to English) we can check for similarities between the original document and the translated back document. One advantage of our graph matching approach is that it is not sensitive to the introduction (or removal) of sentence boundaries between the original and back-translated documents.

### 6.1.1 *Corpus of Translated Documents*

The psychology dataset was collected from psychology literature [**19**] on analogical reasoning and problem solving, consisting of 36 English texts used in several human-subject tests. These texts represent stories containing between 50 and 400 words (average=205) with several being in the form of analogous pairs of stories. A selection of documents (18) from this dataset was translated into different languages and then back-translated to English. Google Translate was used to perform the translations and this translation corpus was created specifically to contribute to the evaluation of Dr Inventor. By varying the difference between English and the target language we aim to evaluate the metrics used by Dr Inventor. Our expectation before undertaking this work was that, as the target language became more distant from English the similarity score between the original and back-translated text should decrease.

The languages chosen were *Irish, Russian, Spanish, French, German, Arabic* and *Amharic*. These languages were selected due to feedback received from native speakers of these languages on Google Translate and as Dr Inventor project members are (mostly native) speakers of these languages. It is expected that Spanish, French and German will be ranked the highest, while Arabic and Amharic will be ranked the lowest. Native speakers of Arabic and Amharic read some sample documents (not part of any Dr Inventor corpus) and they were generally rated as being of poor quality by these speakers. In particular, Amharic was only added to Google Translate in early 2016 and as such, it has not had as long a time to train and refine the translation system. Additionally the languages Russian and Irish were also selected to see if they could be evaluated. Spanish and French are Romance languages with well-developed machine translation systems, so our expectation was that these would produce some of the most faithful translations.

Of course our evaluation will also qualitatively discuss the maturity of Google's translation service for each language.

### 6.1.2 *Translation Quality Estimation*

The best results on the corpus were produced, as expected, for the languages Spanish, French and German. As these are the languages most closely related to English and they are also some of the most widely used and well-developed translation systems. It was decided to use these scores and results to be a baseline for a good translation score. Native English speakers compared the original document with the back-translated document they were generally considered to be fairly accurate re-representations of the original text.

Running the system using the Arabic and Amharic languages also produced the expected results as the scores received were much lower than the "well translated" languages. Native English speakers comparing the original document with the back-translated document agreed that numerous errors did occur. As discussed above, native speakers of these languages did find errors and problems. These were not unexpected due to the dissimilarity in the languages themselves. In particular, in Arabic the word order can be quite different even due to the differences in direction of reading. Additionally, in Arabic, the subject could be dropped from the sentence but still have the same meaning, as the subject is implicitly understood.

Finally, the system was run with our two "testing" languages, Irish and Russian. By using the baseline of the "well translated" languages and the "badly translated" languages, it showed that the Google translate system worked quite well with the Russian and Irish. Their scores were not as high as Spanish, French or German but they were much better performing than Arabic and Amharic.

The box plot below (Figure 7) summarises all the results of this translation evaluation. Overall it showed the Dr Inventor system performed as expected at evaluating the "well-translated" and "badly-translated" languages.

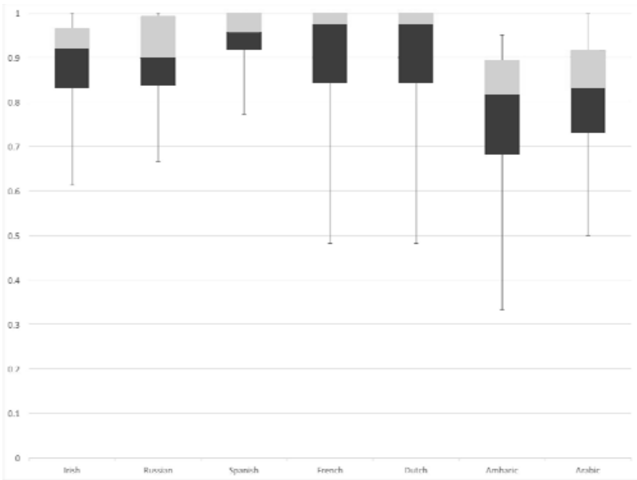


**Figure 7.** Similarity scores for the languages Irish, Russian, Spanish, French, German, Arabic and Amharic

## 6.2 Plagiarism Corpus

A corpus of plagiarised short documents was created [**22**] with the aim that it could be used for the development and evaluation of plagiarism detection tools. The corpus consists of short answers to computer science questions and the plagiarism challenge has been simulated, representing various degrees of plagiarism. Using this corpus we assessed Dr Inventor's ability to detect plagiarism among these documents, i.e. searching for near analogies.

### 6.2.1 *Levels of Plagiarism*

Each answer used a Wikipedia entry as a source text. The corpus has four levels of plagiarism: 1) *near copy*: simply copying text from the entry. 2) *light revision*: basing the answer on the entry but the text could be altered in basic ways. Words could be substituted and paraphrasing could be used. 3) *heavy revision*: again basing the answer on the entry but the text was rephrased using different words and structure. 4) *non-plagiarism*: by using standard learning materials answers were constructed by using the participants own knowledge.

### 6.2.2 *Corpus Contents*

19 participants were asked to answer 5 questions according to the guidelines of the level of plagiarism to be used. 95 answers were generated by these students. Including the original Wikipedia entry 100 documents are contained within the corpus and these documents are passed through the Dr Inventor system to see how it assesses the four different levels of similar contained within this corpus.

### 6.2.3 *Output from the System*

All of the 100 documents were processed by the Dr Inventor system and ROS graphs were created for each of them. The original document was compared against the 4 different plagiarised versions by mapping the respective ROS graphs. The semantic similarity score from Dr Inventor was measured and the following box plot was obtained over the corpus.

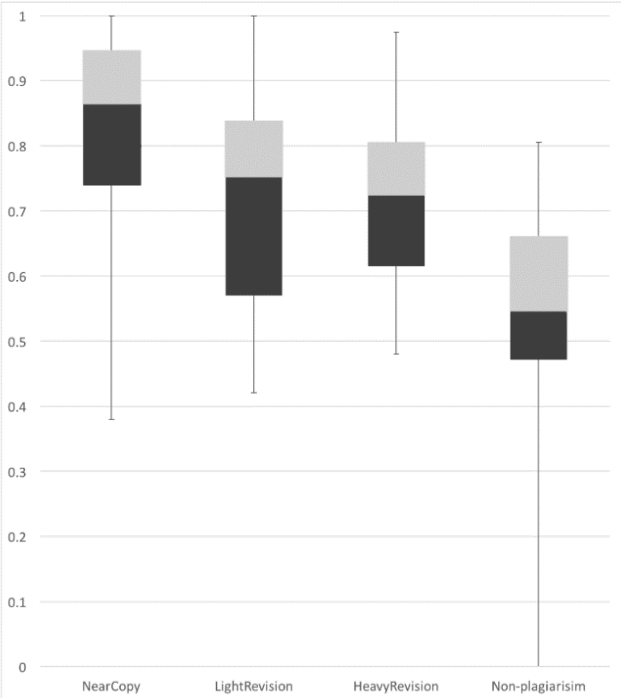


**Figure 8.** Semantic Similarity for Different levels of Plagiarism

This shows that as the amount of plagiarism decreases, the semantic similarity found by Dr Inventor decreases as well. This again was a very pleasing result as it shows that the metrics currently in use by Dr Inventor show a degree of refinement in estimating the similarity between plagiarised versions of documents.

## 6.3 Future Work

Our earlier results show that the existing metrics used by the Dr Inventor system appear to operate effectively, even when there's relatively little semantic distance between the two input documents. This gives us confidence to start exploring its use in dealing with patent applications. Estimating the similarity between patent applications [**23**] is particularly important to Dr Inventor. One current undertaking relates to adapting the parser to correctly handle some of the lexical peculiarities of patents so that they are correctly processed by the parser [**24**].

Some future work is based on the notion that many commercially sensitive patents are written such that they will not be found by existing retrieval tools. This makes the challenge of filing a defence against a new patent application very difficult for the holder of an existing patent. In future work we hope to be able to identify some of these patents.

# 7 Conclusion

We described the Dr Inventor system that identifies figurative and structure-based similarities between text based documents.

The first evaluation used Dr Inventor's metrics to identify two very high quality analogies between publications from the SIGGRAPH conference series. Each was evaluated by a senior researcher in computer graphics and this showed that each comparison was both novel and also represented a reasonable hypothesis that was worthy of their consideration. Both evaluators agreed that each could (at least) be considered as the basis for subsequent research.

The second evaluation of Dr Inventor outlined a translation corpus that was based on English language versions of 18 different texts sourced from various psychological studies on the analogy process. These texts were translated into seven different target languages and then back-translated to English, creating different versions of the source document. Dr Inventor showed that the closest languages produced the best translations as estimated by its metrics. Similarly the newest translation languages which are also the most distant from English produced the lowest results - the two intermediate languages are producing intermediate results. While lacking a certain degree of refinement these results show that Dr Inventor may be usefully used to estimate the quality of "roundtrip" translations.

Dr Inventor and its similarity metrics were also assessed at the task of evaluating a "short answers" plagiarism corpus. This contained short documents in one of three levels of plagiarism, as well as one non-plagiarised version of each document. Again this evaluation showed that Dr Inventor showed a good ability to identify between the different levels of plagiarism within the corpus.

## ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Union Seventh Framework Programme ([FP7/2007-2013]) under grant agreement no 611383.

## REFERENCES

[1] D.P. O'Donoghue, Y. Abgaz, D. Hurley, F. Ronzano, and H. Saggion, "Stimulating and Simulating Creativity with Dr Inventor," in *6th International Conf on Computational Creativity*, Utah, USA, 2015.

[2] M. A. Boden, "Creativity and artificial intelligence," *Artificial Intelligence*, vol. 103, no. 1, pp. 347-356, 1998.

[3] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management*, vol. 24, no. 5, pp. 513-523, 1988.

[4] T K Landauer, P W Foltz, and D Laham, "An Introduction to Latent Semantic Analysis," *Discourse Processes*, vol. 25, pp. 259-284, 1998.

[5] C. Paul, A. Rettinger, A. Mogadala, C. A. Knoblock, and P. Szekely, "Efficient Graph-Based Document Similarity," in *International Semantic Web Conference*, May, 2016, pp. 334-349.

[6] D. Gentner, "Structure-mapping: A theoretical framework for analogy," *Cognitive Science, Volume 7*, pp. 155-170, 1983.

[7] Gilles Fauconnier and Mark Turner, "Conceptual integration networks," *Cognitive science*, vol. 22, no. 2, pp. 133--187, 1998.

[8] T. Veale, D. P. O'Donoghue, and M. T. Keane, "Computation and Blending," *Cognitive Linguistics 11 (3/4)*, pp. 253-281, 2000.

[9] G. A. Miller, "WordNet: A Lexical Database for English.," *Communications of the ACM*, vol. 38, no. 11, pp. 39-41, 1995.

[10] L.P Cordella, P Foggia, C Sansone, and M Vento, "A (sub)graph isomorphism algorithm for matching large graphs," in *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, 2004, pp. 1367-1372.

[11] Dekang Lin, "An Information-Theoretic Definition of Similarity," , San Francisco, CA, USA, 1998.

[12] K. Belhajjame et al., "Workflow-centric research objects: First class citizens in scholarly discourse," in *9th Extended Semantic Web Conference*, Hersonissos, Greece, 2012.

[13] F Ronzano and H Saggion, "Knowledge Extraction and Modeling from Scientific Publications," in *In the Proceedings of the Workshop "Semantics, Analytics, Visualisation: Enhancing Scholarly Data" co-located with the 25th International World Wide Web Conference*, Montreal, Canada, 2016.

[14] A. Constantin, S. Pettifer, and A. Voronkov, "PDFX: fully-automated PDF-to-XML conversion of scientific literature," in *Proceedings of the 2013 ACM symposium on Document engineering*, 2013, pp. 177–180.

[15] B. Bohnet, "Very high accuracy and fast dependency parsing is not a contradiction.," in *Proc. 23rd International Conference on Computational Linguistics*, 2010, pp. 89–97.

[16] Beatriz Fisas, Francesco Ronzano, and Horacio Saggion, "On the Discoursive Structure of Computer Graphics Research Papers," in *The 9th Linguistic Annotation Workshop held in conjuncion with NAACL 2015.*, 2015.

[17] B. Agarwal, S. Poria, N. Mittal, A. Gelbukh, and A. Hussain, "Concept-level sentiment analysis with dependency-based semantic parsing: A novel approach," *Cognitive Computation*, pp. 1-13, 2015.

[18] Keith, J. Holyoak and Paul Thagard, "Analogical mapping by constraint satisfaction," *Cognitive Science*, vol. 13, pp. 295-355, 1989.

[19] M L. Gick and K J. Holyoak, "Analogical problem solving," *Cognitive psychology*, vol. 12, no. 3, pp. 306-355, 1980.

[20] Dedre Gentner and Arthur, B. Markman, "Defining structural similarity," *Journal of Cognitive Science*, vol. 6, pp. 1-20, 2005.

[21] Paul. Jaccard, "Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines," *Bulletin de la Société Vaudoise des Sciences Naturelles*, vol. 37, pp. 241-272, 1901.

[22] P. Clough and M. Stevenson, "Developing A Corpus of Plagiarised Short Answers," *Language Resources and Evaluation: Special Issue on Plagiarism and Authorship Analysis, In Press.*

[23] S. Reardon, "Text-mining offers clues to success," *Nature*, vol. 509, no. 7501, pp. 410-410, 2014.

[24] A. Burga, J. Codina, G. Ferraro, H. Saggion, and L. Wanner, "The challenge of syntactic dependency parsing adaptation for the patent domain.," in *ESSLLI-13 Workshop on Extrinsic Parse Improvement.*, 2013, August.