SIMTEX: An Approach for Detecting and Measuring Textual Similarity based on Discourse and Semantics

Iria da Cunha, Jorge Vivaldi, Juan-Manuel Torres-Moreno, Gerardo Eugenio Sierra-Martinez

SIMTEX: An Approach for Detecting and Measuring Textual Similarity based on Discourse and Semantics

Iria da Cunha[1], Jorge Vivaldi[1], Juan-Manuel Torres-Moreno[2,3], and Gerardo Sierra[2,4]

[1] University Institute for Applied Linguistics (Universitat Pompeu Fabra), Barcelona, Spain

[2] LIA/Agorantic/Université d'Avignon et des Pays de Vaucluse, Avignon, France

[3] École Poytechnique de Montréal, Montréal, (Québec) Canada

[4]Universidad Nacional Autónoma de México/Instituto de Ingeniería, México DF, Mexico

{iria.dacunha,jorge.vivaldi}@upf.edu, juan-manuel.torres@univ-avignon.fr, GSierraM@ii.unam.mx

Abstract. Nowadays automatic systems for detecting and measuring textual similarity are being developed, in order to apply them to different tasks in the field of Natural Language Processing (NLP). Currently, these systems use surface linguistic features or statistical information. Nowadays, few researchers use deep linguistic information. In this work, we present an algorithm for detecting and measuring textual similarity that takes into account information offered by discourse relations of Rhetorical Structure Theory (RST), and lexical-semantic relations included in EuroWordNet. We apply the algorithm, called SIMTEX, to texts written in Spanish, but the methodology is potentially language-independent.

Keywords. Textual similarity, discourse, semantics, paraphrase.

1 Introduction

In the field of Natural Language Processing (NLP), automatic systems for textual similarity detection and measurement are being developed, in order to apply them to different tasks, such as plagiarism detection, question answering, textual entailment, summarization, automatic machine translation evaluation, etc. A lot of research on comparison of long texts has been done, but nowadays a more challenging task is the comparison of short texts, in order to obtain the degree of semantic similarity between them.

As [18] explain, methods for detecting and measuring similarity between short texts can be divided in three groups:

1. Methods that use vector space model [25]. These methods model both texts as a bag of words, and represent them by means of vectors, which are compared by using cosine similarity.
2. Methods that align segments and compute similarity of pairs of words.
3. Methods that use machine learning models combining several measures and lexical, semantic or syntactic features.

In this work, we focus on the second method (see our previous work in [9]). We present an approach for textual similarity detection and measurement that takes into account information offered by discourse relations of Rhetorical Structure Theory (RST) [19], and lexical-semantic relations included in EuroWordNet (EWN)[1]. For the selection and alignment of the segments to be compared, discourse structure is used, while for the selection of

[1]http://www.illc.uva.nl/EuroWordNet

ISSN 1405-5546
DOI: 10.13053/CyS-18-3-2033

the words to be matched EWN is employed. In our approach, one of the main innovations regarding the state of the art is the comparison of discourse segments, instead of complete sentences. Moreover, only discourse segments involving similar RST discourse relations are compared. The pairs of words of these segments are compared by means of EWN, as much work on this field does. However, in this case, the semantic similarity measure [35] is used, which has been already used in other NLP tasks, such as summarization [34] or term extraction [23, 33]. Our approach, called SIMTEX, has been applied to texts written in Spanish, but the methodology is language-independent. The resources necessary to adapt the algorithm to other languages are a discourse parser in the corresponding language, as well as an ontology or lexical database, such as WordNet (or any other resource that allows calculating semantic similarity).

Another contribution of this work is the development of the corpus for the experiments. Developing corpora for the evaluation of systems for detecting textual similarity is a complex and time-consuming task. Regarding corpora in English, two efforts can be pointed out. First, the METER corpus (MEasuringTExtReuse) [10], which is composed by reused texts and their corresponding description regarding their level of relevance and segmentation. Second, the PAN corpus (Plagiarism Corpus) [4], which from 2007 has been a dynamic corpus, since it is used as reference corpus of the most relevant competition on detection of textual similarity and plagiarism. However, to our knowledge, there is no similar resource for Spanish. Therefore, we had to create our own corpus for the experiments. As it can be seen in Section 3, this corpus contains original texts and different related paraphrased texts, which have been written manually. As the authors of [4] state, paraphrases are linguistic expressions having different form but approximately the same meaning, where the form is the lexical or syntactic structure.

In Section 2, related work on textual similarity detection and measurement is reviewed. In Section 3, the theoretical framework and resources used in this work are presented. In Section 4, the design of SIMTEX is explained. In Section 5, an example of application of SIMTEX over the corpus is included. Finally, in Section 6, some conclusions and future work are shown.

2 Related Work

As we have mentioned, textual similarity detection is a challenging task that nowadays is being investigated by several authors. Since some years ago, there has conducted an international competition on semantic textual similarity (see, for example, the last one: "SEM 2013 shared task: Semantic Textual Similarity" [1]). Most of the systems that have been developed for detecting and measuring semantic textual similarity use text pairs to be compared as feature vectors where each feature is a score related to a specific type of similarity. In this section, we do not pretend to list all the work related to semantic textual similarity, but to point out the newest contributions in the field.

The authors of [11] model the task as a Support Vector (SV) regression problem, where a similarity scoring function between pairs of texts is obtained from examples. Semantic relatedness between sentences is modeled in an unsupervised fashion by means of several similarity functions. Each one captures a specific semantic aspect, such as syntactic vs. lexical similarity. The authors of [11], [27] also use dependency parsing, but they model the problem as a combination of kernels [29].

[3] grade pairs of sentences accurately by combining focused measures into a robust measure by means of a log-linear regression model, either based on surface features, on lexical semantics or on Explicit Semantic Analysis. [7] use a SV regression model, combining different text similarity measures that constitute the features. In this case, the measures are simple distances, such as Levenshtein edit distance, cosine or Named Entities overlap, and more complex distances, such as Explicit Semantic Analysis, WordNet-based similarity, IR-based similarity, and a similarity measure based on syntactic dependencies. The authors of [18] also use a SV regression model to combine features. However, they include a semantic word similarity model based on a combination of Latent Semantic Analysis (LSA) [14] and knowledge from WordNet. [28] present a different approach with regard to the methods that use pairwise similarity

features in order to learn a regression model. Their system directly encodes the input texts into syntactic/semantic structures and uses tree kernels for extracting a set of syntactic patterns to learn a similarity score.

In the field of plagiarism, research is also conducted with the aim to obtain methods for detecting similarity between texts, in order to discover if one text is the original and the other one is the copy. For example, [31] uses morphosyntactic information by means of n-grams. This method is useful when some segments in the original text are literally copied in another text. However, in the cases of paraphrased texts (where different words or syntactic structures exist), this method is not enough to obtain accurate results. [15] also use morphosyntactic information but, in order to solve the mentioned limitations, they add semantic information. Specifically, they use WordNet in order to obtain synonyms and hypernyms. In this line, [4] mention that semantic relations, such as synonymy and antonymy, can be used to detect paraphrases. Also, in the field of authorship attribution, methods for detecting textual similarity have been developed. For example, [30] present a system based on syntactic n-grams constructed by following paths in syntactic trees. This method allows bringing syntactic knowledge into machine learning methods. Textual similarity is also important in other NLP tasks, such as machine translation. See, for example, the work in [2], where a semantic feature for statistical machine translation based on Latent Semantic Indexing is proposed.

Textual similarity, paraphrase, plagiarism, reuse of text, etc. are related terms that are difficult to define. In this work, we focus on paraphrase, taking into account that an original text and a paraphrased text can be considered as similar texts. Several authors have offered a definition of paraphrase. As [22] declare, carrying out paraphrases involves using different words and changing syntactic structures from one text to another one, without changing the meaning. This is the reason why lexical paraphrases and morphosyntactic paraphrases exist [6]. Regarding this issue, in the framework of NLP, [32] state that "two units of text are interchangeable if, for the propositions A and B they embody, the truth-set of B is a (not necessarily proper) subset of the truth-set of A".

Several classifications of types of paraphrases exist. [5] offer a classification which includes four classes of paraphrase: Morpholexicon-based changes, Structure-based changes, Semantics-based changes and Miscellaneous changes, which contain 20 types of possible paraphrases. In this work, we use this classification.

3 Theoretical Framework and Resources

In this section the theoretical framework and the resources used in this research are explained. RST is a language-independent theory based on the idea that a text can be segmented into Elementary Discourse Units (EDUs) linked by means of nucleus-satellite or multinuclear rhetorical relations. In the first case, the satellite gives additional information about the other unit (the nucleus), on which it depends (e.g. Cause, Purpose or Result). In the second case, several elements, all nuclei, are connected at the same level, i.e. there are no dependent elements and they all are equally important with regard to the author's intentions (e.g. List, Contrast or Sequence). Discourse parsing includes three stages: discourse segmentation, discourse relations detection and building up rhetorical trees. In this work, a discourse parser for Spanish texts that is integrated in the platform DiZer 2.0 is used [24]. This parser integrates a discourse segmenter [12], a set of linguistic patterns for detecting discourse relations extracted from the RST Spanish Treebank [13] and a probabilistic algorithm for building rhetorical trees [21].

In this work, we also use EWN, which is a multilingual extension of WordNet. In this ontology, the basic semantic unit is the synset (synonymy set), grouping together several words that can be considered synonyms in some contexts. Synsets are linked by means of semantic relations (hyperonym, hyponym, meronym, etc.).

As mentioned in Section 1, in this work we have developed our own corpus in order to exemplify and validate our algorithm. This corpus contains 12 specialized texts from the mathematics domain,

ISSN 1405-5546
DOI: 10.13053/CyS-18-3-2033

divided in 3 original texts (ot), 3 texts with low-level paraphrases (llp), 3 texts with high-level paraphrases (hlp), and 3 texts that are not paraphrases (np), but with similar length, subject and register. The paraphrased texts have been manually built by a team of 3 people, who are Spanish linguists and had a training course on paraphrase techniques, following the classification by [5] mentioned in Section 2. The instructions given to the 3 annotators were:

— For low-level paraphrases: only lexical substitutions can be done in sentences; specifically, only lexical units with the grammatical category of noun, verb, adjective or adverb can be substituted. If a lexical unit can be replaced by a synonym or a hypernym, it is mandatory to replace it. The objective is to change as many units as possible in the sentence. These new units should be searched on general dictionaries, specialized dictionaries and databases from the mathematics domain, or specialized texts on mathematics.

— For high-level paraphrases: the low-level paraphrasis plus additional changes should be done in the text, following the classification by [5], which implies syntactic, semantic, discourse and structural changes. The paraphrased text should be as different as possible from the original text.

See a real example of a sentence from the corpus:

— Original sentence (ot):
Usando algunas propiedades del grupo diédrico, se da una prueba simple de que ciertos arreglos de los números de un famoso rompecabezas (llamado en inglés the fifteen puzzle) no son posibles de realizar. [Using some properties of the dihedral group, it is simply tested that it is not possible for carrying out certain arrangements of the numbers of a famous puzzle (called the fifteen puzzle in English).]

— Low-level paraphrase (llp):
Empleando ciertos rasgos *del grupo diédrico, se* ***otorga*** *una* ***comprobación sencilla*** *de que* ***algunos ajustes*** *de las* ***cifras*** *de un* ***afamado*** *rompecabezas (****denominado*** *en inglés the fifteen puzzle) no son* ***viables*** *de* ***efectuar****.*

— High-level paraphrase (hlp):
Al emplear ciertos rasgos del grupo perteneciente o relativo al ángulo diedro, el cual es cada una de las dos porciones del espacio limitadas por dos semiplanos que parten de una misma recta, se puede comprobar simplemente que algunos ajustes de las cifras de un afamado rompecabezas denominado the fifteen puzzle no son viables de efectuar.

In llp, 13 lexical units have been changed (marked in bold). In hlp, the same units have been replaced and other different changes have been added. For example, the term *grupo diédrico* ("dihedral group") has been replaced by its definition: *grupo perteneciente o relativo al ángulo diedro* ("group belonging or related to the dihedral angle"). Also, the structure of gerund at the beginning of the sentence has been replaced by other equivalent structure in Spanish: *al* ("when") + verb in infinitive. Moreover, among other changes, brackets have been eliminated.

The aim of this paper is not to explain the details of the methodology for building this corpus, but to exemplify our algorithm for detecting and measuring textual similarity, and to obtain preliminary results to validate it and continue with the implementation and further experiments. We are conscious that the size of the corpus is limited, but it should be taken into account that paraphrases of the original texts have been done manually, which is a very time-consuming and difficult task. In the future, we plan to increase the corpus size including more texts from other domains. This corpus will be available on line for research purposes.

At the moment, it would not be possible to carry out our experiments for English, since we need a dataset containing original texts and paraphrased texts, both annotated with RST discourse structure. On the one hand, we could obtain original annotated texts from the RST Discourse Treebank [8], the biggest corpus for English including texts annotated with RST discourse structure; however, this

corpus does not contain paraphrased texts. On the other hand, we could access some datasets for English built in the framework of textual similarity detection (such as [11] and [12], mentioned in Section 1); however, these datasets are not annotated with discourse structure and, nowadays, to our knowledge, the only discourse full parser for English [20] is not available to the scientific community. Therefore, it is not possible to obtain texts discourse-annotated automatically for English at the moment. For other languages, several research groups work on they own discourse parsers; we highlight the full discourse parser available for Portuguese [26].

4 Design of the Algorithm

The design of the algorithm includes three modules, which are explained in this section. The algorithm has been implemented in Perl.

MODULE 1: DISCOURSE COMPARISON

In the first place, the discourse parser is used to obtain RST discourse trees of the two texts (A and B) to be compared. The output of the parser includes two files for each discourse tree: a file containing the detected discourse segments and a file containing the discourse relations and structure of the text in parenthetical format.

The following example shows an original text from our corpus (called text A) and the two files obtained automatically with the discourse parser:

— **Text A**

El objetivo de este trabajo es dar una justificación rigurosa del método general de prueba conocido como inducción y del método general de definición conocido como recursión, que ocurren frecuentemente tanto en lógica matemática como en otras ramas de la matemática. Es muy importante tener claras algunas hipótesis bajo las cuales estos métodos son válidos. En lo que sigue se usará la teoría de los conjuntos de un modo intuitivo, así como ejemplos que se suponen conocidos, de aritmética, álgebra y lógica. [The goal of this work is to give a rigorous justification of the general test method known as induction and the definition general method known as recursion, which are frequently used in mathematical logic as well as in other branches of mathematics. It is important to know some hypotheses under which these methods are valid. In what follows, the set theory will be used intuitively, and also known examples of arithmetic, algebra and logic will be employed.]

— **FILE 1: Discourse segments**

1: El(el)$_D$ objetivo(objetivo)$_N$ de(de)$_S$ este(este)$_D$ trabajo(trabajo)$_N$ es(ser)$_V$ dar(dar)$_V$ una(uno)$_D$ justificación(justificación)$_N$ rigurosa(riguroso)$_A$ de(de)$_S$ el(el)$_D$ método(método)$_N$ general(general)$_A$ de(de)$_S$ prueba(prueba)$_N$ conocido(conocer)$_V$ como(como)$_C$ inducción(inducción)$_N$ y(y)$_C$ de(de)$_S$ el(el)$_D$ método(método)$_N$ general(general)$_A$ de(de)$_S$ definición(definición)$_N$ conocido(conocer)$_V$ como(como)$_C$ recursión(recursión)$_N$,(,)$_F$ [s] que(que)$_P$ ocurren(ocurrir)$_V$ frecuentemente(frecuentemente)$_R$ tanto(tanto)$_R$ en(en)$_S$ lógica(lógica)$_N$ matemática(matemático)$_A$ como(como)$_C$ en(en)$_S$ otras(otro)$_D$ ramas(rama)$_N$ de(de)$_S$ la(el)$_D$ matemática(matemática)$_N$.(.)$_F$ [s]

2: Es(ser)$_V$ muy(muy)$_R$ importante(importante)$_A$ tener(tener)$_V$ claras(clara)$_N$ algunas(alguno)$_D$ hipótesis(hipótesis)$_N$ bajo(bajo)$_S$ las(el)$_D$ cuales(cual)$_P$ estos(este)$_D$ métodos(método)$_N$ son(ser)$_V$ válidos(válido)$_A$.(.)$_F$ [s]

3: En(en)$_S$ lo(el)$_D$ que(que)$_P$ sigue(seguir)$_V$ se(se)$_P$ usará(usar)$_V$ la(el)$_D$ teoría(teoría)$_N$ de(de)$_S$ los(el)$_D$ conjuntos(conjunto)$_N$ de(de)$_S$ un(uno)$_D$ modo(modo)$_N$ intuitivo(intuitivo)$_A$,(,)$_F$ [s] así_como(así_como)$_C$ ejemplos(ejemplo)$_N$ que(que)$_P$ se(se)$_P$ suponen(suponer)$_V$ conocidos(conocer)$_V$,(,)$_F$ [s] de(de)$_S$ aritmética(aritmética)$_N$,(,)$_F$ [s] álgebra(álgebra)$_N$ y(y)$_C$ lógica(lógica)$_N$.(.)$_F$ [s] [p]

— **FILE 2: Discourse structure**

elaboration(n('means(n(1), s(2))'), s(3))

In the second place, our algorithm compares the discourse relations included in the parenthetical discourse structures of both texts (A and B), using file 2 as input. It detects if there

are identical discourse relations between discourse structures of both texts. Then, it calculates a score of discourse similarity by taking into account the amount of identical relations detected between both texts, as well as the differences between both texts. In such a way, it does not only takes into account the similarity but the dissimilarity among the parenthetical discourse relations of both texts. Thus, the algorithm calculates the difference of the intersection of identical relations between text A and text B, minus the edit distance (DR(A,B)) [17] of the parenthetical discourse structures between texts A and B. We normalize the score by using the sum of the total relations included in both texts (see Formula 1).

$$Sim^D = \frac{2 \times \mathrm{R}(A) \cap \mathrm{R}(B) - \mathrm{DR}(A,B)}{|\mathrm{R}(A)| + |\mathrm{R}(B)|} \quad (1)$$

Let's see an example. Figures 1 and 2 include the discourse structures of text A and B:

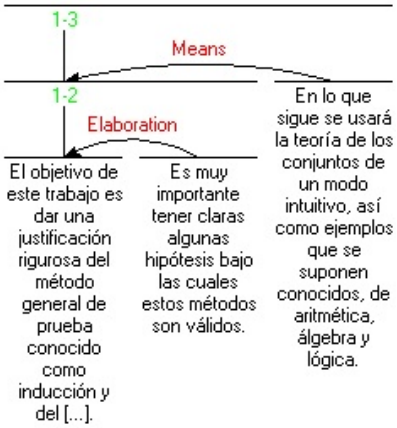

Fig. 1. RST discourse structure of text A

— **Text B**

El propósito de esta investigación es otorgar argumentos precisos del procedimiento de inducción, el cual es un método generalmente utilizado para probar o demostrar que una afirmación dada es verdadera para todos los números naturales; y del procedimiento de recursión, que a su vez, se utiliza para determinar el siguiente término de una secuencia utilizando uno o más de los términos anteriores. Tanto el método de inducción como el de recursión acontecen a menudo tanto en lógica matemática como en otras especialidades de la matemática. Pero el método de inducción como el método de recursión se pueden realizar bajo ciertas posibilidades las cuales se deben tener claras. Se utilizará de forma intuitiva la teoría de los conjuntos, la cual es estudio de la estructura y tamano de conjuntos desde el punto de vista de los axiomas aplicados. Además se ejemplificará con reconocidos modelos de aritmética, álgebra y lógica. [The purpose of this research is to give precise arguments about the procedure of induction, which is a method generally used to prove or to demonstrate that a given statement is true for all the natural numbers; and about the procedure of recursion, which, in turn, is used to determine the next term of a sequence using one or more of the previous terms. Both induction and recursion methods appear in mathematical logic as well as in other fields of mathematics. But the induction method as well as the recursion method can be carried out under certain possibilities that must be known. The set theory will be used intuitively, which means the study of structure and size of sets from the point of view of applied axioms. Moreover, it will be exemplified with well-known models of arithmetic, algebra and logic.]

Text A contains 2 relations (Elaboration, Means), while text B contains 4 relations (Elaboration, Antithesis, Means, Elaboration). Therefore, the relation of Elaboration has 2 intersections between A and B; the relation of Means has 1 intersection, and the relation of Antithesis has not any intersection (that is, 0).

In the third place, the difference between texts A and B is calculated through the edit distance of their parenthetical discourse relations. The costs of insertion and deletion are 1, while for substitution the cost is 2. For the example above, the cost to

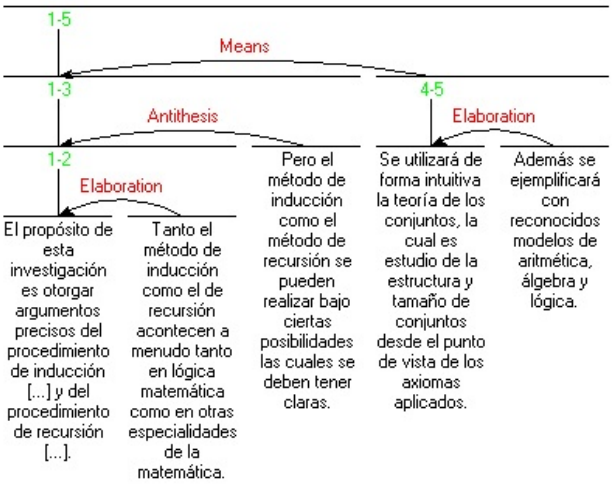

Fig. 2. RST discourse structure of text B

transform (Elaboration, Antithesis, Means, Elaboration) into (Elaboration, Means) is 2.

Thus, by applying Formula 1, the obtained score of discourse similarity is 0.67, as shown in the Formula 2.

$$Sim^D = \frac{2 \times (2+1) - 2}{2+4} = 0.67 \qquad (2)$$

If this score is higher or equal than 0, the algorithm extracts pairs of discourse segments including identical relations between texts A and B, and also the main nuclei of both discourse trees, by using file 1 (that is, the file containing the discourse segments detected by the parser). In this example, the pairs of extracted segments are:

A: *El objetivo de este trabajo es dar una justificación rigurosa del método general de prueba conocido como inducción y del método general de definición conocido como recursión, que ocurren frecuentemente tanto en lógica matemática como en otras ramas de la matemática.*
B: *El propósito de esta investigación es otorgar argumentos precisos del procedimiento de inducción, el cual es un método generalmente utilizado para probar o demostrar que una afirmación dada es verdadera para todos los números naturales; y del procedimiento de recursión, que a su vez, se utiliza para determinar el siguiente término de una secuencia utilizando uno o más de los términos anteriores.*

A: *Es muy importante tener claras algunas hipótesis bajo las cuales estos métodos son válidos.*
B: *Tanto el método de inducción como el de recursión acontecen a menudo tanto en lógica matemática como en otras especialidades de la matemática.*

A: *Es muy importante tener claras algunas hipótesis bajo las cuales estos métodos son válidos.*
B: *Además se ejemplificará con reconocidos modelos de aritmética, álgebra y lógica.*

A: *En lo que sigue se usará la teoría de los conjuntos de un modo intuitivo, así como ejemplos que suponen conocidos, de aritmética, álgebra y lógica.*
B: *Se utilizará de forma intuitiva la teoría de los conjuntos, la cual es estudio de la estructura y tamaño de conjuntos desde el punto de vista de los axiomas aplicados.*

These segments will be used in Module 2, in order to calculate the semantic similarity between them. By contrast, if the score is below 0, semantic similarity will not be calculated.

In the future and after further experiments, a threshold higher than 0 will be determined in this module. This threshold will be used to determine if the algorithm should continue by applying Module 2, or if there is not any discourse similarity between text A and B and, therefore, the algorithm should stop the process after applying Module 1.

MODULE 2: SEMANTIC COMPARISON

In this module, in the first place, the extracted discourse segments are lemmatized. Nouns are extracted, taking into account that this kind of lexical units usually includes the most representative information of the text, specially in texts from specialized domains. In the second place, the algorithm calculates similarity between pairs of lexical units in the discourse segments of texts A and B

including identical relations. In order to calculate similarity between these pairs of lexical units, we apply Formula 3, which uses information obtained from the hyperonimical paths for each synset (sy) in EWN.

$$Sim(sy_1, sy_2) = \frac{2 + \#\text{CommonNodes}(sy_1, sy_2)}{\text{Depth}(sy_1) + \text{Depth}(sy_2)} \quad (3)$$

This similarity measure is based on [35]. They obtain a similarity value by combining the depth of two concepts and the depth of the least common subsumed node in a "IS-A" hierarchical concept net like EWN. This similarity measure takes into account two basic ideas: a) the shorter the distance between two nodes is, the higher their similarity is, and b) the higher the number of common nodes is (therefore lower in the hierarchy), the higher their similarity is. In practice, the similarity between two terms like "argument" and "definition" is calculated as Figure 3 shows.

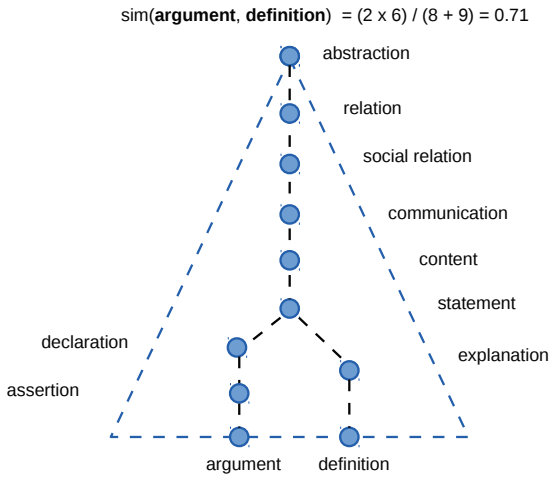

Fig. 3. Example of semantic similarity calculation

For example, let's consider the following two segments:

— Segment A (Elaboration): *Es muy importante tener claras algunas hipótesis bajo las cuales estos métodos son válidos.*

— Segment B (Elaboration): *Además se ejemplificará con reconocidos modelos de aritmética, álgebra y lógica.*

The nouns *hipótesis* ("hypothesis") and *método* ("method") are extracted from segment A. The nouns *modelo* ("model"), *aritmética* ("arithmetic"), *álgebra* ("algebra") and *lógica* ("logic") are extracted from segment B. Therefore, the algorithm will compare the following units:

hipótesis - modelo
hipótesis - aritmética
hipótesis - álgebra
hipótesis - lógica

método - modelo
método - aritmética
método - álgebra
método - lógica

At this stage, the methodology includes three steps. First, each lexical pair comparison obtains a semantic similarity score between 0 and 1. Second, all the scores of each segment are added, in order to obtain a single semantic similarity score for each pair of discourse segments. Third, the scores of all discourse segments of each text (A and B) are added, in order to obtain the final semantic similarity score between the two original texts. The score is normalized between 0 and 1.

MODULE 3: COMPUTING THE FINAL TEXTUAL SIMILARITY SCORE

In this module, discourse and semantic scores are combined. In our current work, the discourse similarity score obtained accounts for 30% of the score, while the semantic similarity score accounts for 70% of the score. The final score is normalized between 0 and 1. As shown in Section 6, in the future we plan to perform experiments with different percentages, but in the current research we have used these values taking into account that, once identical discourse relations are detected, the semantic score is crucial in order to detect and measure textual similarity.

5 Example of Application

We have applied SIMTEX to our Spanish corpus on mathematics, as stated in Section 3. In Table 1[2] discourse similarity is included, considering the similarity and the difference between each pair of texts. The maximum score is given to those texts presenting just the same parenthetical relations, i.e., equal similarity but no dissimilarity. On the other side, the worst score is for the non-paraphrased texts.

Table 1. Discourse similarity

A	B	|R(A)|	|R(B)|	R(A)∩R(B)	DR(A,B)	Sim(A,B)^D
5ot	5llp	1	1	1	0	+1.00
5ot	5hlp	1	1	1	0	+1.00
5ot	5np	1	1	0	2	-1.00
7ot	7llp	2	2	2	0	+1.00
7ot	7hlp	2	4	3	2	+0.67
7ot	7np	2	2	2	2	+0.50
9ot	9llp	2	2	2	0	+1.00
9ot	9hlp	2	2	1	2	0.00
9ot	9np	2	2	0	4	-1.00

Using Formula 3, we calculate semantic similarity for nouns among all the pairs of discourse segments, except for those texts with a threshold lower than 0. Table 2 shows discourse and semantic similarities (Sim^D and Sim^S), as well as the normalized final score (Sim^T). As shown in Table 2, the Sim^T score obtained between original texts and non-paraphrased texts is low, as expected. By contrast, the Sim^T score between original texts and paraphrased texts is, in most cases, higher. The score of low-level paraphrases is the highest in all cases.

In order to compare our results, we have defined a baseline similarity Sim^B as follows:

$$Sim^B = 2 \times \frac{|\text{bigrams}(A) \cap \text{bigrams}(B)|}{|\text{bigrams}(A)| + |\text{bigrams}(B)|} \quad (4)$$

In our work, the baseline similarity is calculted by the following two different strategies: the first one uses a stoplist (BL1), and the second one uses all words of documents (BL2) before computing Formula 4. In Table 3, similarity results of both baselines are included. Numbers in bold indicate the best results.

[2]Texts are identified by a number plus a code of text type, as indicated in Section 3.

Table 2. Final textual similarity score

A	B	Sim(A,B)^D	Sim(A,B)^S	Sim(A,B)^T
5ot	5llp	+1.000	0.417	**0.892**
5ot	5hlp	+1.000	0.455	**0.919**
5ot	5np	-1.000	0.000	**0.000**
7ot	7llp	+1.000	0.438	**0.907**
7ot	7hlp	+0.670	0.352	**0.747**
7ot	7np	+0.500	0.237	0.616
9ot	9llp	+1.000	0.365	**0.855**
9ot	9hlp	0.000	0.201	**0.440**
9ot	9np	-1.000	0.000	**0.000**

Table 3. Textual vs. Baseline similarity

A	B	Sim(A,B)^T	BL1(A,B)	BL2(A,B)
5ot	5llp	**0.892**	0.250	0.344
5ot	5hlp	**0.919**	0.115	0.161
5ot	5np	**0.000**	**0.000**	0.036
7ot	7llp	**0.907**	0.250	0.447
7ot	7hlp	**0.747**	0.134	0.176
7ot	7np	0.616	**0.000**	0.067
9ot	9llp	**0.855**	0.146	0.288
9ot	9hlp	**0.440**	0.045	0.106
9ot	9np	**0.000**	0.025	0.041

These numbers show that our method allows to discriminate between paraphrased and non-paraphrased texts, while performance of both baseline strategies is worst. However, for 7ot and 7np texts, our method reports a high similarity value (0.616) that is incorrect. This value is obtained because, in this case, the discourse similarity between both texts is high.

6 Conclusions and Future Work

In this work, we have presented an algorithm for detecting and measuring textual similarity that takes into account information offered by discourse

relations of RST, and lexical-semantic relations included in EWN. We have applied the algorithm to Spanish texts, but the methodology is language-independent.

Although the amount of texts included in the corpus is not big, as we have mentioned, the main goal of our research has been to show the algorithm and to obtain preliminary results in order to validate it. These preliminary results indicate that the performance of the algorithm is promising.

As future work, we plan to increase the corpus size and extend our experiments to other grammatical categories (verbs, adjectives and adverbs). Also, we will do further experiments for optimizing the threshold included in Module 1. Moreover, we will do experiments with other lexical databases. Although WordNet is largely employed in NLP applications, it is still far from covering all existing words and senses. In our case, the Spanish EWN version used includes about 25,000 synsets (corresponding to 50,000 variants). Thus, the performance of our similarity algorithm can be affected by this reason.

In future, we plan to carry out experiments with the Multilingual Central Repository (MCR)[3] (an improved and expanded version based on both EWN and WordNet 3.0) [16], and also with the structure of pages and categories of Wikipedia. Finally, we will integrate the different modules into a complete single robust automatic system.

Acknowledgments

We acknowledge the Mexico's National Council of Science and Technology (Conacyt) grant number 178248 and Project UNAM-DGAPA-PAPIIT number IN400312. We also acknowledge the support of the Spanish projects RICOTERM 4 (FFI2010-21365-C03-01) and APLE 2 (FFI2012-37260), a Juan de la Cierva grant (JCI-2011-09665) and an Ibero-America Young Teachers and Researchers Santander Grant 2013.

[3] http://adimen.si.ehu.es/web/MCR

References

1. **Agirre, E., Cer, D., Diab, M., González-Agirre, A., & Guo, W.** (**2013**). SEM 2013 shared task: Semantic Textual Similarity. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*, volume 1. Association for Computational Linguistics, Atlanta, Georgia, USA, 32–43.
2. **Banchs, R. & Costa-jussá, M.** (**2011**). A semantic feature for statistical machine translation. In *Proceedings of SSST-5, Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation. ACL HLT 2011*. Portland, Oregon, 126–134.
3. **Bär, D., Biemann, C., Gurevych, I., & Zesch, T.** (**2012**). Ukp: Computing semantic textual similarity by combining multiple content similarity measures. In *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, volume 2. Association for Computational Linguistics, Montreal, Canada, 435–440.
4. **Barrón-Cedeño, A., Potthast, M., Rosso, P., Stein, B., & Eiselt, A.** (**2010**). Corpus and Evaluation Measures for Automatic Plagiarism Detection. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association, Valletta, Malta, 771–774.
5. **Barrón-Cedeño, A., Vila, M., Martí, M., & Rosso, P.** (**2013**). Plagiarism Meets Paraphrasing: Insights for the Next Generation in Automatic Plagiarism Detection. *Computational Linguistics*, 39(4), 917–947.
6. **Barzilay, R. & McKeown, K. R.** (**2001**). Extracting Paraphrases from a Parallel Corpus. In *Proceedings of the 39th Annual Meeting of the ACL*. Association for Computational Linguistics, Toulouse, France, 50–57.
7. **Buscaldi, D., Le Roux, J., Garcia Flores, J., & Popescu, A.** (**2013**). LIPN-CORE: Semantic Text Similarity using n-grams, WordNet, Syntactic Analysis, ESA and Information Retrieval based Features. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*, volume 1. Association for Computational Linguistics, Atlanta, Georgia, USA, 162–168.
8. **Carlson, L., Marcu, D., & Okurowski, M. E.** (**2002**). RST Discourse Treebank. In *Pennsylvania: Linguistic Data Consortium*. Pennsylvania.
9. **Castro Rolón, B., Sierra, G., Torres-Moreno, J.-M., & da Cunha, I.** (**2011**). El discurso y la

semántica como recursos para la detección de similitud textual. In *Proceedings of the III RST Meeting (8th Brazilian Symposium in Information and Human Language Technology, STIL 2011)*. Brazilian Computer Society, Cuiabá, Brasil.

10. **Clough, P., Gaizauskas, R., & Piao, S.** (**2002**). Building and annotating a corpus for the study of journalist text reuse. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, volume 5. Las Palmas, Canary Islands, Spain, 1678–1691.
11. **Croce, D., Annesi, P., Storch, V., & Basili, R.** (**2012**). Unitor: Combining semantic text similarity functions through sv regression. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, volume 1. Association for Computational Linguistics, Montreal, Canada, 597–602.
12. **da Cunha, I., SanJuan, E., Torres-Moreno, J.-M., Lloberes, M., & Castellón, I.** (**2012**). DiSeg 1.0: The First System for Spanish Discourse Segmentation. *Expert Systems with Applications*, 39(2), 1671–1678.
13. **da Cunha, I., Torres-Moreno, J.-M., & Sierra, G.** (**2011**). On the Development of the RST Spanish Treebank. In *Proceedings of the 5th Linguistic Annotation Workshop*. Association for Computational Linguistics, Portland, Oregon, USA, 1–10.
14. **Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R.** (**1990**). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6), 391–407.
15. **Dolan, B., Quirk, C., & Brockett, C.** (**2004**). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING'04. Association for Computational Linguistics, Geneva, Switzerland, 1–7.
16. **Gonzalez-Agirre, A., Laparra, E., & Rigau, G.** (**2012**). Multilingual central repository version 3.0. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey.
17. **Levenshtein, V.** (**1966**). Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, 10(8), 707–710.
18. **Lushan, H., Kashyap, A., Finin, T., Mayfield, J., & Weese, J.** (**2013**). UMBC_EBIQUITY-CORE: Semantic Textual Similarity Systems. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*, volume 1. Association for Computational Linguistics, Atlanta, Georgia, USA, 44–52.
19. **Mann, W. C. & Thompson, S. A.** (**1988**). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3), 243–281.
20. **Marcu, D.** (**2000**). The Rhetorical Parsing of Unrestricted Texts: A Surface-based Approach. *Computational Linguistics*, 26(3), 395–448.
21. **Marcu, D.** (**2000**). *The Theory and Practice of Discourse Parsing Summarization*. The MIT Press, Cambridge, MA, USA. ISBN 0262133725.
22. **Maurer, H., Kappe, F., & Zaka, B.** (**2006**). Plagiarism - A survey. *Journal of Universal Computer Science*, 12(8), 1050–1084.
23. **Maynard, D.** (**1999**). *Term recognition using combined knowledge sources*. Ph.D. thesis, Manchester Metropolitan University, Faculty of Science and Engineering.
24. **Maziero, E., Pardo, T., da Cunha, I., Torres-Moreno, J.-M., & SanJuan, E.** (**2011**). DiZer 2.0-An Adaptable On-line Discourse Parser. In *Proceedings of the III RST Meeting (8th Brazilian Symposium in Information and Human Language Technology)*. 50–57.
25. **Meadow, C. T.** (**1992**). *Text Information Retrieval Systems*. Academic Press, Inc., Orlando, FL, USA.
26. **Pardo, T. & Nunes, M.** (**2008**). On the development and evaluation of a brazilian portuguese discourse parser. *Journal of Theoretical and Applied Computing*, 15(2), 43–64.
27. **Polajnar, T., Rimell, L., & Kiela, D.** (**2013**). UCAM-CORE: Incorporating structured distributional similarity into STS. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*, volume 1. Association for Computational Linguistics, Atlanta, Georgia, USA, 85–89.
28. **Severyn, A., Nicosia, M., & Moschitti, A.** (**2013**). iKernels-Core: Tree Kernel Learning for Textual Similarity. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*, volume 1. Association for Computational Linguistics, Atlanta, Georgia, USA, 53–58.
29. **Shawe-Taylor, J. & Cristianini, N.** (**2004**). *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA. ISBN 0521813972.

30. **Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., & Chanona-Hernandez, L.** (**2014**). Syntactic N-grams as machine learning features for natural language processing. *Expert Systems with Applications*, 41(3), 853–860.
31. **Spassova, M. S.** (**2009**). *El potencial discriminatorio de las secuencias de categorías gramaticales en la atribución forense de autoría de textos en español*. Ph.D. thesis, IULA, Universitat Pompeu Fabra, Barcelona.
32. **Vila, M., Martí, A., & Rodríguez, H.** (**2011**). Paraphrase Concept and Typology. A Linguistically Based and Computationally Oriented Approach. *Procesamiento del Lenguaje Natural*, 46, 83–90.
33. **Vivaldi, J.** (**2001**). *Extracción de candidatos a términos mediante combinación de estrategias heterogéneas*. Ph.D. thesis, IULA, Universitat Pompeu Fabra, Barcelona.
34. **Vivaldi, J., da Cunha, I., Torres-Moreno, J. M., & Velázquez-Morales, P.** (**2010**). Automatic summarization using terminological and semantic resources. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association, Valletta, Malta, 3105–3112.
35. **Wu, Z. & Palmer, M.** (**1994**). Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the ACL*. Association for Computational Linguistics, Las Cruces, New Mexico, USA, 133–138.

Iria da Cunha holds a Ph.D. in Applied Linguistics from the Universitat Pompeu Fabra (UPF) in Barcelona. Nowadays, she holds a Juan de la Cierva research contract in the framework of the group IULATERM (Lexicon and Technology), from the University Institute for Applied Linguistics (IULA). Also, she is associated lecturer at the Faculty of Translation and Interpretation of UPF. Her main research lines are discourse parsing, automatic summarization, specialized discourse analysis and terminology.

Jorge Vivaldi Palatresi obtained his Ph.D. degree from the Polytechnical University of Catalonia with a dissertation focused on extracting terms from written texts in the biomedical area. Currently, he is a researcher at the University Institute for Applied Linguistics, Universitat Pompeu Fabra in Barcelona, where he is responsible for the coordination of several projects dealing with corpus processing and information extraction. His areas of interest are mainly related to natural language processing, both resources compilation and tools development.

Juan-Manuel Torres-Moreno obtained his Ph.D. degree in Computer Science (Neural Networks) from Institut National Polytechnique de Grenoble and his HDR degree from Laboratoire Informatique d'Avignon (LIA). Nowadays he is full Professor at the LIA (Universite d'Avignon et des Pays de Vaucluse), where he is responsible of the NLP team (TALNE), and for the coordination of projects with information extraction. His areas of interest are mainly related to NLP, information extraction and automatic text summarization.

Gerardo Sierra Martínez is a National Reseacher of Mexico. He leads the Grupo de Ingeniería Lingüística at the Instituto de Ingeniería of the Universidad Nacional Autónoma de México (UNAM). He holds a Ph.D. in Computational Linguistics from the University of Manchester, Institute of Science and Technology (UMIST), UK. His research interest is focused on language engineering and includes computational lexicography, concept extraction, corpus linguistics, text mining and forensic linguistics.

Article received on 20/01/2014; accepted on 21/03/2014.