Assessing Comparable Corpora through Distributional Similarity Measures

Hernani Costa
LEXYTRAD, University of Malaga, Spain
hercos@uma.es

Abstract

Describing, comparing and evaluating corpora are key issues in corpus-based translation and corpus linguistics for which there is still a notable lack of standards. Bearing this in mind, this paper aims at investigating the use of textual distributional similarity measures in the context of comparable corpora. More precisely, we address the issue of measuring the relatedness between documents by extracting and measuring their common content. For this purpose, we designed and applied a methodology that exploits available natural language processing technology with statistical methods. Our findings showed that using a list of common entities and a simple, yet robust and high performance set of distributional similarity measures was enough to describe and assess the degree of relatedness between the documents in a comparable corpus.

1 Introduction

The use of comparable corpora has been considered an essential resource in several research domains such as Natural Language Processing (NLP), terminology, language teaching, and automatic and assisted translation, amongst others. Nevertheless, an inherent problem to those who deal with comparable corpora in a daily basis is the uncertainty about the data they are dealing with. Indeed, little work has been done on automatically characterising such linguistic resources and attempting a meaningful description of their content is often a perilous task (Corpas Pastor and Seghiri, 2009). Usually, a corpus is given a short description such as “casual speech transcripts” or “tourism specialised comparable corpus”. However, such tags will be of little use to those users seeking for a representative and/or high quality domain-specific corpora. Apart from the usual description that comes along with the corpus, like number of documents, tokens, types, source(s), creation date, policies of usage, etc., nothing is said about how similar the documents are. As a result, most of the resources at our disposal are built and shared without deep analysis of their content, and those who use them blindly trust on the people’s or research group’s name behind their compilation process, without knowing nothing about the relatedness quality of the corpus.

Bearing this in mind, in this work we try to fill this void by taking advantage of several textual distributional similarity measures presented in the literature. First, we selected a specialised corpus about tourism and beauty domain that was manually compiled by researchers in the area of translation and interpreting studies. Then, we designed and applied a methodology that exploits available NLP technology with statistical methods to assess how the documents correlate with each other in the corpus. Our assumption is that the amount of information contained in a document can be evaluated via summing the amount of information contained in the member words. For this purpose, a list of common entities was used as a unit of measurement capable of identifying the amount of information shared between the documents. Our assumption is that this approach will allow us not only to compute the relatedness between documents, but also to describe and characterise the corpus itself.

The remainder of the paper is structured as follows. Section 2 introduces some fundamental concepts related to distributional similarity measures, i.e. explains the theoretical foundations, related work and the distributional similarity exploited in this experiment. Then, Section 3 presents the corpus used in this work. After applying the methodology described in Section 4, Section 5 presents and discusses the obtained results in detail. Finally, Section 6 presents the final remarks and highlights our future plans for this work.

2 Distributional Similarity Measures

Information Retrieval (IR) (Singhal, 2001) is the task of locating specific information within a collection of documents or other natural language resources according to some request. In this field, we can find a large number of statistical methods based on words and their (co-)occurrence. Essentially, it involves finding the most frequently used words and treating the rate of usage of each word in a given text as a quantitative attribute. Then, these words serve as features for a given statistical method. Following Harris' distributional hypothesis (Harris, 1970), which assumes that similar words tend to occur in similar contexts, these statistical methods are suitable, for instance to find similar sentences based on the words they contain (Costa et al., 2015a) and automatically extract or validate semantic entities from corpora (Costa et al., 2010; Costa, 2010; Costa et al., 2011). To this end, it is assumed that the amount of information contained in a document could be evaluated by summing the amount of information contained in the document words. And, the amount of information conveyed by a word can be represented by means of the weight assigned to it (Salton and Buckley, 1988). Accordingly, we took advantage of two IR measures commonly used in the literature, the Spearman's Rank Correlation Coefficient (SCC) and the Chi-Square (χ^2) to compute the similarity between two documents written in the same language (see section 2.1 and 2.2). Both measures are particularly useful for this task because they are independent of text size (mostly because both use a list of the common entities), and they are language-independent.

The Spearman's Rank Correlation Coefficient (SCC) distributional measure has been shown effective on determining similarity between sentences, documents and even on corpora of varying sizes (Kilgarriff, 2001; Costa et al., 2015a). It is particularly useful, for instance to measure the textual similarity between two documents because it is easy to compute and is independent of text size as it can directly compare ranked lists for large and small texts.

The χ^2 similarity measure has also shown its robustness and high performance. By way of example, χ^2 have been used to analyse the conversation component of the British National Corpus (Rayson et al., 1997), to compare corpora (Kilgarriff, 2001), and to identify topic related clusters in imperfect transcribed documents (Ibrahimov et al., 2002). It is a simple statistic measure that permits to assess if relationships between two variables in a sample are due to chance or the relationship is systematic.

For all these reasons, distributional similarity measures in general and SCC and χ^2 in particular have a wide range of applicabilities (cf. Kilgarriff (2001) and Costa et al. (2015a)). Indeed, this work aims at proving that these simple, yet robust and high-performance measures allow to describe the relatedness between documents in specialised corpora.

2.1 Spearman's Rank Correlation Coefficient (SCC)

In this work, the SCC is adopted and calculated as in Kilgarriff (2001). Firstly, a list of the common entities[1] L between two documents d_l and d_m is compiled, where $L_{d_l,d_m} \subseteq (d_l \cap d_m)$. It is possible to use the top n most common entities or all common entities between two documents, where n corresponds to the total number of common entities considered $|L|$, i.e. $\{n|n \in \mathbb{N}^0, n \leq |L|\}$ – in this work we use all the common words for each document pair, i.e. $n = |L|$. Then, for each document the list of common entities (e.g. L_{d_l} and L_{d_m}) is ranked by frequency in an ascending order ($R_{L_{d_l}}$ and $R_{L_{d_m}}$), where the entity with lowest frequency receives the numerical raking position 1 and the entity with highest frequency receives the numerical raking position n. In the case of ties in rank, where more than one entity in a document occurs with the same frequency, the average of the ranks is assigned to the tying entities. For instance, if the entities e_a, e_b and e_c had the same frequency and ranked in the 6^{th}, 7^{th} and 8^{th} position, all three entities would be assigned the same rank of $\frac{6+7+8}{3}$ =7. Finally, for each common entity $\{e_1, ..., e_n\} \in L$, the difference in the rank orders for the entity in each document is computed,

[1] In this work, the term 'entity' refers to "single words", which can be a token, a lemma or a stemm.

and then normalised as a sum of the square of these differences $\left(\sum_{i=1}^{n} s_i^2\right)$. The final SCC equation is presented in expression 1, where $\{SCC|SCC \in \mathbb{R}, -1 \geq SCC \leq 1\}$.

By a way of example let e_x be a common entity (i.e. $\{e_x\} \in L$) and $R_{L_{d_l}} = \{1\#e_{n_{d_l}}, 2\#e_{n-1_{d_l}}, ..., n\#e_{1_{d_l}}\}$ and $R_{L_{d_m}} = \{1\#e_{n_{d_m}}, 2\#e_{n-1_{d_m}}, ..., n\#e_{1_{d_m}}\}$ the resulting ranked list of common words for d_l and d_m, respectively. Supposing that e_x is the $3\#e_{n-2_{d_l}}$ and $1\#e_{n_{d_m}}$, i.e. e_x is in the 3^{rd} position in $R_{L_{d_l}}$ and in the 1^{st} position in $R_{L_{d_m}}$, s would be computed as $s_{e_x}^2 = (3-1)^2$ and the result would be 4. Then, this process is repeated for the remain $n-1$ entities and the resulted SCC score will be seen as the similarity value between d_l and d_m.

$$SCC(d_i, d_j) = 1 - \frac{6 * \sum_{i=1}^{n} s_i^2}{n^3 - n} \tag{1}$$

2.2 Chi-Square (χ^2)

The Chi-square (χ^2) measure also uses a list of common words (L). Similarly to SCC, it is also possible to use the top n most common entities or all common entities between two documents, and again in this work we use all the common words for each document pair, i.e. $n = |L|$. The number of occurrences of a common words in L that would be expected in each document is calculated from the frequency lists. If the size of the document d_l and d_m are N_l and N_m and the entity e_i has the following observed frequencies $O(e_i, d_l)$ and $O(e_i, d_m)$, then the expected values are $e_{i_{d_l}} = \frac{N_l*(O(e_i,d_l)+O(e_i,d_m))}{N_l+N_m}$ and $e_{i_{d_m}} = \frac{N_m*(O(e_i,d_l)+O(e_i,d_m))}{N_l+N_m}$. Equation 2 presents the χ^2 formula, where O is the observed frequency and E the expected frequency. The resulted χ^2 score should be interpreted as the interdocument distance between two documents. It is also important to mention that $\{\chi^2|\chi^2 \in \mathbb{R}, 1 \geq \chi^2 < \infty\}$, which means that as more unrelated the common words in L are, the lower the χ^2 score will be.

$$\chi^2 = \sum \frac{(O-E)^2}{E} \tag{2}$$

Suppose that we have two common entities e_i and e_j between two documents d_l and d_m (i.e. $L = \{e_i, e_j\}$). Table 1 shows a contingency table example. This table contains: i) the observed frequencies (O); ii) the totals in the margins; iii) and the expected frequencies (E), which are obtained by applying the following formula: $\frac{column\ total}{N} * row\_total$, e.g. $E(e_i, d_l) = \frac{14}{26} * 15 = 8.08$. After writing down the expected frequencies in the table, we are ready to calculate the χ^2 score (see Equation 3).

	d_l	d_m	*Total*
e_i	O=11 E=8.08	O=4 E=6.92	15
e_j	O=3 E=5.92	O=8 E=5.08	11
Total	14	12	26

Table 1: Example of a contigency table.

$$\chi^2 = \frac{(11-8.08)^2}{8.08} + \frac{(3-5.92)^2}{5.92} + \frac{(4-6.92)^2}{6.92} + \frac{(8-5.08)^2}{5.08} = 5.41 \tag{3}$$

3 The INTELITERM Corpus

The INTELITERM[2] corpus is a comparable corpus composed of documents collected from the Internet. Designed to be a specialised comparable corpus, this corpus was manually compiled by researchers

[2] http://www.lexytrad.es/proyectos.html

with the purpose of building a representative corpus for the Tourism and Beauty domain. It contains documents in four different languages (English, Spanish, German and Italian). Some of the texts are translations of each other, yet the majority is composed of original texts. The INTELITERM comparable corpus is composed of several subcorpora, divided by the language and further for each language there are translated and original texts (which will be hereafter referred as language_totd and language_to, respectively). In this work, we used half of the corpus, i.e. all the original and translated documents in English and Spanish (en_to, en_totd, es_to and es_totd, respectively). All the information about these subcorpora is presented in Table 2. In detail, this table shows: the number of documents (nDocs); the number of types (types); the number of tokens (tokens); and the ratio of types per tokens ($\frac{types}{tokens}$) per subcorpus. These values were obtained using the corpus analysis toolkit for concordancing and text analysis software Antconc 3.4.3 (Anthony, 2014).

	nDocs	types	tokens	$\frac{types}{tokens}$	description
en_to	151	11,6k	508,9k	0.023	original
en_totd	61	6,9k	88,5k	0.078	translated
es_to	225	12,6k	253,4k	0.049	original
es_totd	27	3,4k	19,7k	0.174	translated

Table 2: Statistical information about the various subcorpus.

4 Methodology

This section not only describes the methodology used to calculate the similarity between documents using Distributional Similarity Measures (DSMs), but also presents all the tools, libraries and frameworks employed by our system to perform this experiment.

1) **Data Preprocessing**: firstly all the documents within the corpus were processed with the OpenNLP[3] Sentence Detector and Tokeniser. Then, the annotation process was done with the TT4J[4] library, which is a Java wrapper around the popular TreeTagger (Schmid, 1995) – a tool specifically designed to annotate text with part-of-speech and lemma information. Regarding the stemming, we used the Porter stemmer algorithm provided by the Snowball[5] library. A method to remove punctuation and special characters within the words was also implemented. Finally, in order to get rid of the noise, a stopword list[6] was compiled to filter out the most frequent words in the corpus. Once a document is computed and the sentences are tokenised, lemmatised and stemmed, our system creates a new output file with all this new information, i.e. the new document contains: the original, the tokenised, the lemmatised and the stemmed text. Using the stopword list mentioned above a Boolean vector describing if the entity is a stopword or not is also added. This way, the system will be able to use only the tokens, lemmas and stems that are not stopwords.

2) **Identifying the list of common entities between documents**: in order to identify a list of common entities (L), a co-occurrence matrix was built for each pair of documents. Only those that have at least one occurrence in both documents are considered. As required by the DSMs (see section 2), their frequency in both documents is also stored within this matrix ($L_{d_l,d_m} = \{e_i,(f(e_i,d_l),f(e_i,d_m)); e_j,(f(e_j,d_l),f(e_j,d_m)); ...; e_n,(f(e_n,d_l),f(e_n,d_m))\}$). With the purpose of analysing and comparing the performance of different DSMs, three different lists were created to be used as input features: the first one using common tokens, another using common lemmas and the third one using common stems.

3) **Computing the similarity between documents**: the similarity between documents was calculated by applying three different DSMs ($DSMs = \{DSM_{NCE}, DSM_{SCC}, DSM_{\chi^2}\}$, where $_{NCE}$, $_{SCC}$ and

[3] https://opennlp.apache.org

[4] http://reckart.github.io/tt4j/

[5] http://snowball.tartarus.org

[6] Freely available to download through the following URL https://github.com/hpcosta/stopwords.

χ^2 means Number of Common Entities, Spearman's Rank Correlation Coefficient and Chi-Square, respectively), each one calculated using three different input features (list of common tokens, lemmas and stems).

4) **Computing the document final score**: the document final score $DSM(d_l)$ is the mean of the similarity scores of the document with all the documents in the collection of documents, i.e. $DSM(d_l) = \frac{\sum_{i=1}^{n-1} DSM_i(d_l,d_i)}{n-1}$, where n corresponds to the total number of documents in the collection and $DSM_i(d_l, d_i)$ the resulted similarity score between the document d_l with all the documents in the collection.

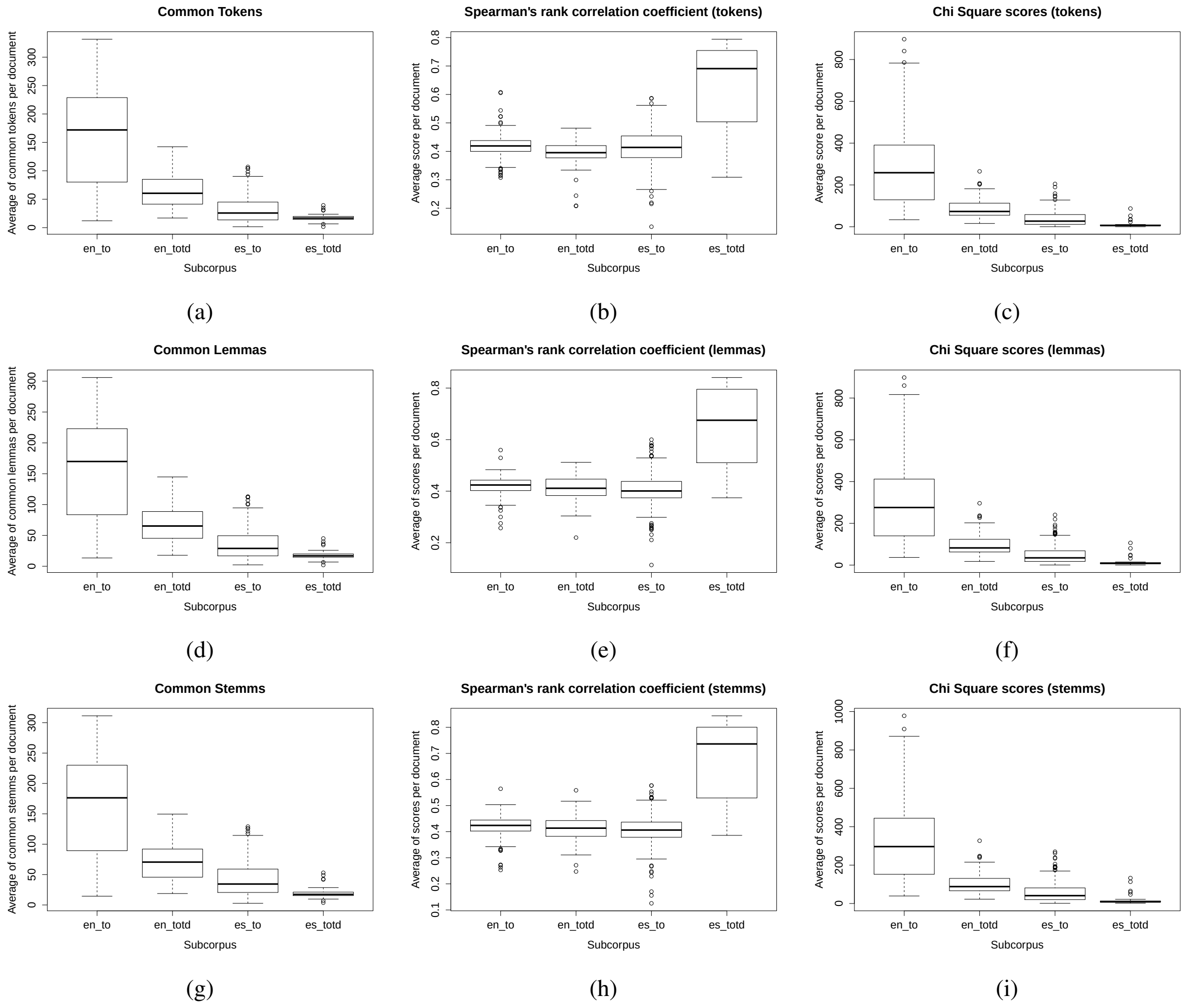
5 Results and Analysis

In order to describe the corpus in hand, we applied three different Distributional Similarity Measures (DSMs): the Number of Common Entities (NCE), the Spearman's Rank Correlation Coefficient (SCC) and the Chi-Square (χ^2). As a input feature to the DSMs, three different types of entities (tokens, lemmas and stems) were used. Table 3 shows the Number of Common Tokens (NCT) between document on average (av), the SCC and the χ^2 scores along with the associated standard deviations (σ) per measure and subcorpus. Figure 1 presents the resulted average scores per document in a box plot format for all the combinations DSM vs. feature. Each box plot displays the full range of variation (from min to max), the likely range of variation (the interquartile range), the median, and the high maximums and low minimums (also know as outliers). It is important to mention that for this experiment we did not use a sample, but instead the entire corpus in its original size and form, which means that all obtained results and made observations came from the entire population, in this case the various INTELITERM English (en_to and en_totd) and Spanish (es_to and es_totd) subcorpora.

		NCT	**SCC**	χ^2
en_to	av	163.70	0.42	279.39
	σ	83.87	0.05	177.45
en_totd	av	67.54	0.39	90.38
	σ	35.35	0.05	53.25
es_to	av	31.97	0.41	40.92
	σ	23.48	0.07	38.21
es_totd	av	17.93	0.63	13.40
	σ	8.46	0.14	18.95

Table 3: Average and standard deviation of common tokens scores between document per subcorpus.

The first observation we can make from Figure 1 is that the distributions between the features are quite similar (see for instance Figures 1a, 1d and 1g). This means that it is possible to achieve acceptable results only using raw words (i.e. tokens). Stems and lemmas require more processing power and time to be used as features – especially lemmas due to the Part-of-Speech (POS) tagger dependency and time consuming process implied. In general, we can say that the scores for each subcorpus is symmetric (roughly the same on each side when cut down the middle), which means that the data is normally distributed. There are some exceptions such as the SCC and χ^2 average scores for the es_totd and for the en_to, respectively, which we will discuss later in this section. Another interesting observation is related with the high NCE (see Table 3 and Figures 1a, 1d and 1g) in original documents (en_to and es_to) when compared with documents translated from other languages (en_totd and es_totd, respectively). For example, the subcorpus en_to (which contains original documents) has 163.70 common tokens per document on average (av) with a standard deviation (σ) of 83.87 and the subcorpus en_totd (which contains translated documents) only has 67.54 common tokens per document on average with a σ=35.35 (Table 3). The same observation can be made between the es_to and the es_totd subcorpus (see Figure 1a and Table 3). This fact could happen because these documents are collections of translated documents collected from the Internet, and thus translated from different translator, which implies that different translators use different vocabulary and consequently lower the NCE between the documents will be.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 1: INTELITERM Subcorpus: average scores per document.

Although the Number of Common Tokens (NCT) per document on average is higher for the en_to subcorpus, the interquartile range (IQR) is larger than for the other subcorpora (see Table 3 and Figure 1a), which means that the middle 50% of the data is more distributed and thus the average of NCT per document is more variable. Moreover, longest whiskers (the lines extending vertically from the box) in Figure 1a also indicates variability outside the upper and lower quartiles. Therefore, we can say that en_to has a wide type of documents and consequently some of them are only roughly correlated to the rest of the subcorpus. Nevertheless, the data is skewed right, which means that the majority is strongly similar, i.e. the documents have a high degree of relatedness between each other. This idea can be sustained by the positive average SCC scores presented in Figure 1b and the set of outliers found above the upper whisker. Moreover, the average of 0.42 SCC score and σ=0.05 also implies a strong correlation between the documents in the en_to subcorpus. Likewise, the longest whisker outside the upper quartile and the skewed left χ^2 scores also indicate relatedness between the documents.

Regarding the en_totd subcorpus, the NCT, the SCC and the χ^2 scores (Figures 1a, 1b and 1c) and the average of 90.38 common tokens per document and σ=53.25 (Table 3) suggest that the data is either normally distributed (Figure 1b) or skewed left (Figures 1a and 1c). Considering this results, we can conclude that the documents are highly related.

From all the subcorpora, the es_to subcorpus is the biggest one with 225 documents, 12606 types, 253412 tokens (Table 2). Nevertheless, Table 3 and Figure 1a reveal a lower NCT compared with en_to and the en_totd subcorpora. A theoretical explanation for this phenomenon is that Spanish has richer morphology compared to English. Therefore, due to bigger number of inflection forms per lemma, there is a larger number of tokens and consequently less common tokens per document in Spanish. When analysing Figures 1a and 1c, the box plots for the es_to subcorpus look similar to the en_totd when shifted up. Except for the longest whisker observed in Figure 1b, the SCC scores also show similar distributions, averages and standard deviations (see Table 3).

As we can see in Figures 1a, 1b and 1c, the average scores per document for es_totd are slightly different from the other box plots. Apart from the low NCT per document, the χ^2 standard deviation higher than its average (18.95 and 13.40, respectively), the SCC variability inside and outside the IQR indicates some inconsistency in the data. This instability can be explained by the subcorpus size, i.e. the small number of documents (27) and by the the low number of types and tokens (3433 and 19736, respectively) and its 0.174 $\frac{types}{tokens}$ ratio. As mentioned by Baker (2006:52), the $\frac{types}{tokens}$ ratio tends to be useful when looking at relatively small documents, and in this specific case this subcorpus only has on average 731 tokens ($\frac{19736}{27} \approx 731$) and 127 types per document ($\frac{3433}{27} \approx 127$), which makes it an excellent test case. When compared with the low ratios from the other subcorpora (see Table 2), – even for this specilised subcorpus – this one can be considered high. If by on one hand, a low ratio can indicate a great number of repetitions (the same word occurring again and again) likely indicating a relatively narrow range of subjects. On the other hand, a high ratio suggests that a more diverse form of language is employed, which can also explain the low NCT and χ^2 scores for this subcorpus in hand. Despite the high SCC, the data is asymmetric and variable (large IQR). This happens because most of the common entities have a low frequency in the documents and consequently they will rank close together in the ranking lists, which results in high SCC scores mostly because of the resulted high value in the numerator (see Equation 1).

To sum up, we can state from the statistical and theoretical evidences that the en_to, the en_totd and the es_to subcorpora look like they assemble highly correlated documents. We can not say the same for the es_totd subcorpus. Due to the small number of documents and scarceness of evidences we can only not reject the idea that this subcorpus is composed of similar documents.

6 Conclusions and Future Work

In this paper we presented and studied various Distributional Similarity Measures (DSMs) for the purpose of describing specialised comparable corpora. As input for these DSMs, we used three different features (lists of common tokens, lemmas and stems). In the end, we conclude that for the data in hand these features had similar performance for all the tested DSMs. In fact, our findings show that instead of using common lemmas or stems, which require external libraries, processing power and time, a simple list of common tokens was enough to describe our data. Moreover, we proved that the corpus used in this experiment is composed of highly correlated documents. The high number of entities shared by its documents, the positive average scores obtained with the SCC measure and their χ^2 scores sustain our claim.

In the immediate future, we intend not only to perform more experiments with these DSMs by adding noisy documents (i.e. out of topic documents) to the corpus and analyse the DSMs performance, but also merge the translated documents from other languages with original ones and prove that translated documents decrease the general relatedness score. Moreover, it is our intention to do the same experiment with other languages, like Italian and German. Apart from that, we also want to test other DSMs, such as Jaccard, Lin and PMI and compare their performance.

Furthermore, these DSMs can be seen as a suitable tool to rank documents by their similarities, which we believe that will be a handy feature to those who manually or semi-automatically compile corpora mined from the Internet. It will allow them to filter out documents with a low level of relatedness when compared with the rest of the documents in the corpus. Indeed, it is our intention to integrate this methodology in the iCorpora application, an ongoing project that aims to design and develop a robust and agile web-based application capable of semi-automatically compile multilingual comparable and parallel corpora (Costa et al., 2014; Costa et al., 2015c; Costa et al., 2015b).

Acknowledgements

Hernani Costa is supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement n^{o} 317471. Also, the research reported in this work has been partially carried out in the framework of the Educational Innovation Project TRADICOR (PIE 13-054, 2014-2015); the R&D project INTELITERM (ref. n^{o} FFI2012-38881, 2012-2015); and the R&D Project for Excellence TERMITUR (ref. n^{o} HUM2754, 2014-2017). I would like to thank Prof. Gloria Corpas Pastor, Prof. Ruslan Mitkov and Dr. Miriam Seghiri for their valuable comments and suggestions to improve the paper.

References

Laurence Anthony. 2014. AntConc (Version 3.4.3) Machintosh OS X. Waseda University. Tokyo, Japan. Available from `http://www.laurenceanthony.net`.

Paul Baker. 2006. *Using Corpora in Discourse Analysis*. Bloomsbury Discourse. Bloomsbury Academic.

Gloria Corpas Pastor and Míriam Seghiri. 2009. Virtual Corpora as Documentation Resources: Translating Travel Insurance Documents (English-Spanish). In A. Beeby, P.R. Inés, and P. Sánchez-Gijón, editors, *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate*, Benjamins translation library, chapter 5, pages 75–107. John Benjamins Publishing Company.

Hernani Costa, Hugo Gonçalo Oliveira, and Paulo Gomes. 2010. The Impact of Distributional Metrics in the Quality of Relational Triples. In *19th European Conf. on Artificial Intelligence, Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, ECAI'10, pages 23–29, Lisbon, Portugal, August.

Hernani Costa, Hugo Gonçalo Oliveira, and Paulo Gomes. 2011. Using the Web to Validate Lexico-Semantic Relations. In *15th Portuguese Conf. on Artificial Intelligence*, volume 7026 of *EPIA'11*, pages 597–609, Lisbon, Portugal, October. Springer.

Hernani Costa, Gloria Corpas Pastor, and Miriam Seghiri. 2014. iCompileCorpora: A Web-based Application to Semi-automatically Compile Multilingual Comparable Corpora. In *Translating and the Computer 36 - AsLing*, London, UK, November.

Hernani Costa, Hanna Béchara, Shiva Taslimipoor, Rohit Gupta, Constantin Orasan, Gloria Corpas Pastor, and Ruslan Mitkov. 2015a. MiniExperts: An SVM approach for Measuring Semantic Textual Similarity. In *9th Int. Workshop on Semantic Evaluation*, SemEval'15, pages 96–101, Denver, Colorado, June. ACL.

Hernani Costa, Gloria Corpas Pastor, Ruslan Mitkov, and Miriam Seghiri. 2015b. $(In\ press)$ Towards a Web-based Tool to Semi-automatically Compile, Manage and Explore Comparable and Parallel Corpora. In *7th Int. Conf. of the Iberian Association of Translation and Interpreting Studies*, AIETI, Malaga, Spain.

Hernani Costa, Gloria Corpas Pastor, Miriam Seghiri, and Ruslan Mitkov. 2015c. iCorpora: Compiling, Managing and Exploring Multilingual Data. In *7th Int. Conf. of the Iberian Association of Translation and Interpreting Studies*, AIETI, pages 74–76, Malaga, Spain, January.

Hernani Costa. 2010. Automatic Extraction and Validation of Lexical Ontologies from text. Master's thesis, University of Coimbra, Faculty of Sciences and Technology, Department of Informatics Engineering, Coimbra, Portugal, September.

Zelig Harris. 1970. Distributional Structure. In *Papers in Structural and Transformational Linguistics*, pages 775–794. D. Reidel Publishing Company, Dordrecht, Holland.

Oktay Ibrahimov, Ishwar Sethi, and Nevenka Dimitrova. 2002. The Performance Analysis of a Chi-square Similarity Measure for Topic Related Clustering of Noisy Transcripts. In *16th Int. Conf. on Pattern Recognition*, volume 4, pages 285–288. IEEE Computer Society.

Adam Kilgarriff. 2001. Comparing Corpora. *Int. Journal of Corpus Linguistics*, 6(1):97–133.

Paul Rayson, Geoffrey Leech, and Mary Hodges. 1997. Social Differentiation in the Use of English Vocabulary: Some Analyses of the Conversational Component of the British National Corpus. *Int. Journal of Corpus Linguistics*, 2(1):133–152.

Gerard Salton and Christopher Buckley. 1988. Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, 24(5):513–523.

Helmut Schmid. 1995. Improvements In Part-of-Speech Tagging With an Application To German. In *ACL SIGDAT-Workshop*, pages 47–50, Dublin, Ireland.

Amit Singhal. 2001. Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(4):35–42.