Measuring the Relatedness between Documents in Comparable Corpora

Hernani Costa[a], **Gloria Corpas Pastor**[a] **and Ruslan Mitkov**[b]
[a]LEXYTRAD, University of Malaga, Spain
[b]RIILP, University of Wolverhampton, UK
{hercos,gcorpas}@uma.es, r.mitkov@wlv.ac.uk

Abstract

This paper aims at investigating the use of textual distributional similarity measures in the context of comparable corpora. We address the issue of measuring the relatedness between documents by extracting, measuring and ranking 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 set of distributional similarity measures was enough to describe and assess the degree of relatedness between the documents. Moreover, our method has demonstrated high performance in the task of filtering out documents with a low level of relatedness. By a way of example, one of the measures got 100%, 100%, 95% and 90% precision when injected 5%, 10%, 15% and 20% of noise, respectively.

1 Introduction

Comparable corpora[1] can be considered an important resource for several research areas such as Natural Language Processing (NLP), terminology, language teaching, and automatic and assisted translation, amongst other related areas. 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 semi- or 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". Yet, 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 or how to retrieve the most related ones. 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 documents. Although some tasks require documents with a high degree of relatedness between each other, the literature is scarce on this matter.

Accordingly, this work explores this niche by taking advantage of several textual Distributional Similarity Measures (DSMs) presented in the literature. Firstly, 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

[1]I.e. corpora that include similar types of original texts in one or more language using the same design criteria (cf. (EAGLES, 1996; Corpas Pastor, 2001)).

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 hypothesis is that this approach will allow us to: compute the relatedness between documents; describe and characterise the corpus itself; and to rank the documents by their degree of relatedness. In order to evaluate how the DSMs perform the task of ranking documents based on their similarity and filter out the unrelated ones, we introduced noisy documents, i.e. out-of-domain documents to the corpus in hand.

The remainder of the paper is structured as follows. Section 2 introduces some fundamental concepts related with DSMs, i.e. explains the theoretical foundations, related work and the DSMs exploited in this experiment. Then, Section 3 presents the corpora 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 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. This field is rich in statistical methods that use words and their (co-)occurrence to retrieve documents or sentences from large data sets. In simple words, these IR methods aim to find the most frequently used words and treat 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., 2015) 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). Having this in mind, 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 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 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., 2015; Costa, 2015). It is particularly useful, for instance to measure the textual similarity between 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 both documents and corpora (Kilgarriff, 2001; Costa, 2015), 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.

Bearing this in mind, distributional similarity measures in general and SCC and χ^2 in particular have a wide range of applicabilities (Kilgarriff, 2001; Costa et al., 2015; Costa, 2015). 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 and to rank them according to their similarity.

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[2] 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

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

common entities between two documents, where n corresponds to the total number of common entities considered $|L|$, i.e. $\{n|n \in N^0, n \leq |L|\}$ – in this work we use all the common entities 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. 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, 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 $\{SCC|SCC \in R, -1 \geq SCC \leq 1\}$.

$$SCC(d_l, d_m) = 1 - \frac{6 * \sum_{i=1}^{n} s_i^2}{n^3 - n} \quad (1)$$

2.2 Chi-Square (χ^2)

The Chi-square (χ^2) measure also uses a list of common entities (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, we use all the common entities for each document pair, i.e. $n = |L|$. The number of occurrences of a common entity 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 $\{\chi^2|\chi^2 \in R, 1 \geq \chi^2 < \infty\}$, which means that as more unrelated the common entities in L are, the lower the χ^2 score will be.

$$\chi^2(d_l, d_m) = \sum \frac{(O - E)^2}{E} \quad (2)$$

3 Corpora

INTELITERM[3] is a specialised comparable corpus composed of documents collected from the Internet. It was manually compiled by researchers with the purpose of building a representative corpus (Biber, 1988, p.246) for the Tourism and Beauty domain. It contains documents in four different languages (English, Spanish, Italian and German). Some of the texts are translations of each other (parallel), yet the majority is composed of original texts. The corpus is composed of several subcorpora, divided by the language and further for each language there are translated and original texts. For the purpose of this work, only original documents in English, Spanish and Italian were used, which for now on will be referred as int_en, int_es, int_it, respectively.

In order to analyse how the DSMs perform the task of ranking documents based on their similarity and filter out the unrelated ones, it is necessary to introduce noisy documents, i.e. out-of-domain documents to the various subcorpora. To do that, we chose the well-known Europarl[4] corpus (Koehn, 2005), a parallel corpus composed by proceedings of the European Parliament. As mentioned further in section 5.2, we added different amounts of noise to the various subcorpora, more precisely 5%, 10%, 15% and 20%. These noisy documents were randomly selected from the "one per day" Europarl v.7 for the three working languages: English, Spanish and Italian (eur_en, eur_es, eur_it, respectively).

	nDocs	types	tokens	$\frac{types}{tokens}$
int_en	151	11,6k	496,2k	0.023
eur_en	30	3.4k	29,8k	0.116
int_es	224	13,2k	207,3k	0.063
eur_es	44	5,6k	43,5k	0.129
int_it	150	19,9k	386,2k	0.052
eur_it	30	4,7k	29,6k	0.159

Table 1: Statistical information per subcorpora.

All the statistical information about both the INTELITERM subcorpora and the set of 20% of noisy documents, randomly selected for each working language, are presented in Table 1. In detail, this Table shows: the number of documents

[3] http://www.lexytrad.es/proyectos.html
[4] http://www.statmt.org/europarl/

(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 Antconc 3.4.3 (Anthony, 2014) software, a corpus analysis toolkit for concordancing and text analysis.

4 Methodology

This section describes the methodology employed to calculate and rank documents based on their similarity using Distributional Similarity Measures (DSMs). All the tools, libraries and frameworks used for the purpose in hand are also pointed out.

1) **Data Preprocessing**: firstly all the INTELITERM documents were processed with the OpenNLP[5] Sentence Detector and Tokeniser. Then, the annotation process was done with the TT4J[6] 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[7] 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[8] 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. a new document containing: 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 to the document. 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 (from now on we will use the acronym NCE), 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 ($L_{d_l,d_m} = \{e_i, (f(e_i, d_l), f(e_i, d_m)); e_j, (f(e_j, d_l), f(e_j, d_m)); ...; e_n, (f(e_n, d_l), f(e_n, d_m))\}$, where f represents the frequency of an entity in a document). 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 the Number of Common Tokens (NCT), another using the Number of Common Lemmas (NCL) and the third one using the Number of Common Stems (NCS).

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 $_{\chi^2}$ refer to Number of Common Entities, Spearman's Rank Correlation Coefficient and Chi-Square, respectively), each one calculated using three different input features (NCT, NCL and NCS).

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) **Ranking documents**: finally, the documents were ranked in a descending order according to their DSMs scores (i.e. NCE, SCC or χ^2).

5 Results and Analysis

This experiment is divided into two parts. In the first part (section 5.1), we describe the corpus in hand by applying three different Distributional Similarity Measures (DSMs): the Number of Common Entities (NCE), the Spearman's Rank

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

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

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

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

Correlation Coefficient (SCC) and the Chi-Square (χ^2). As a input feature to the DSMs, three different lists of entities were used, i.e. the Number of Common Tokens (NCT), the Number of Common Lemmas (NCL) and the Number of Common Stems (NCS). By a way of example, Table 2 shows the NCT between documents, the SCC and the χ^2 scores and averages (av) 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 or IQR), the median, and the high maximums and low minimums (also know as outliers). It is important to mention that for the first part of this experiment (section 5.1) we did not use a sample, but instead the entire INTELITERM subcorpora in their original size and form, which means that all obtained results and made observations came from the entire population, in this case the English (int_en), Spanish (int_es) and Italian (int_it) subcorpora (for more details about the subcorpora see section 3). Regarding the second part of this experiment, we used the same subcorpora, but an additional percentage of documents was added to them in order to test how the DSMs perform the task of filtering out these noisy documents, i.e. out-of-domain documents (see 5.2). In detail, Figure 2 shows how the average scores decrease when injecting noisy documents and Table 3 presents how the DSMs performed when that noise was injected.

5.1 Describing the Corpus

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 tagger dependency and time consuming process implied. In general, we can say that the scores for each subcorpus are symmetric (roughly the same on each side when cut down the middle), which means that the data is normally distributed. There are some exception that we will discuss along this section. Another interesting observation is related with the high Number of Common Tokens (NCT) in English (int_en) when compared with Italian and Spanish (int_it and int_es, respectively), see Table 2 and Figure 1a. Later in this section, we will try to explain this phenomenon.

SubC.	Stats	NCT	SCC	χ^2
int_en	av	163.70	0.42	279.39
	σ	83.87	0.05	177.45
int_es	av	31.97	0.41	40.92
	σ	23.48	0.07	38.21
int_it	av	101.08	0.39	201.97
	σ	55.71	0.05	144.68

Table 2: Average and standard deviation of common tokens scores between documents per subcorpus.

Although the NCT per document on average is higher for the int_en subcorpus, the interquartile range (IQR) is larger than for the other subcorpora (see Table 2 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 int_en 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 left and the longest whisker outside the upper quartile indicates that the majority of the data is strongly similar, i.e. the documents have a high degree of relatedness between each other. This idea can be sustained not only by the positive average SCC scores, but also by the set of outliers above the upper whisker in Figure 1b. The average of 0.42 SCC score and σ=0.05 also implies a strong correlation between the documents in the int_en subcorpus (Table 2). Likewise, the longest whisker and the set of outliers outside the upper quartile in the χ^2 scores also indicate a high relatedness between the documents.

Regarding the int_it subcorpus, the SCC and the χ^2 scores (Figures 1b and 1c) and the average of 101.08 common tokens per document and σ=55.71 (Figure 1a and Table 2) suggest that the data is normally distributed (Figure 1b) and highly

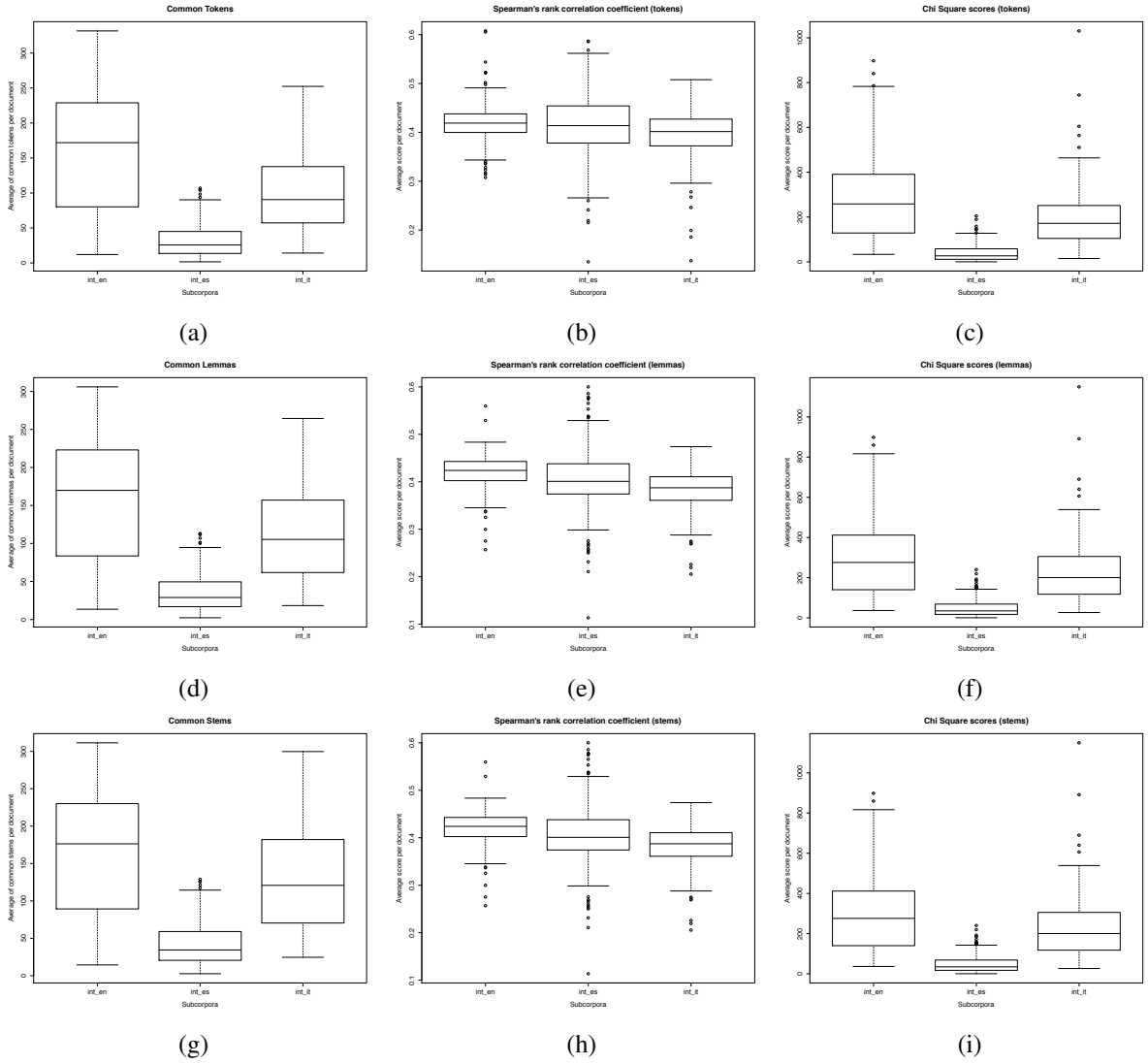

Figure 1: INTELITERM: average scores between documents per subcorpus.

correlated. Although this subcorpus got lower average scores for all the DSMs when compared to the English subcorpus, Table 2, Figure 1a, 1b and 1c show that the average scores and the range of variation are quite similar to the English subcorpus. Therefore, we can conclude that the documents inside the Italian subcorpus are highly related between each other.

From the three subcorpora, the int_es subcorpus is the biggest one with 224 documents (Table 1). Nevertheless, the average scores per document are slightly different from the other box plots (see Figures 1a, 1b and 1c). The χ^2 standard deviation practically equal to its average (38.21 and 40.92, respectively) and the SCC variability inside and outside the IQR indicates some inconsistency in the data. Moreover, Table 2 and Figure 1a reveal a lower NCT compared with int_en and the int_it subcorpora.

The subcorpus int_en has 163 common tokens per document on average with a σ=83, and the subcorpora int_it and int_es only have 101 and 31 common tokens per document on average with a σ=55 and σ=23, respectively (Table 2, NCT column). This means that the int_it and int_es subcorpora are composed of documents with a lower level of relatedness when compared with the English one. This fact could happen because Italian and Spanish have a 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. Another explanation could come from the fact

that the tourism and beauty services are more developed in Italy and Spain than in the UK and therefore there are more variety on the vocabulary used as well as in the services offered. Indeed, Table 1 offers some evidences about the employed vocabulary. The English subcorpus has a lower number of types and a higher number of tokens (11,6k and 496,2k, respectively) when compared with the Italian (19,9k types and 386,2k tokens) and Spanish subcorpora (13,2k types and 207,3k tokens). The high difference on the average of common tokens per document between Spanish and the other two languages can also be related with the marketing strategies used to advertise tourism and beauty services, which is somehow hard to confirm. Despite that our method is able to catch the lexical level of similarity between the documents, the semantic level is not taken into account, i.e. does not consider synonyms as similar words for example, and consequently would result on slightly different similarity scores (again, another explanation difficult to confirm).

To conclude, we can state from the statistical and theoretical evidences that the int_en and the int_it subcorpora look like they assemble highly correlated documents. We can not say the same for the int_es subcorpus. Due to the scarceness of evidences, we can only not reject the idea that this subcorpus is composed of similar documents. Nevertheless, as we will see in the next section, the fact that int_es is composed by low related documents (according to our findings) will affect the ranking task.

5.2 Measuring DSMs Performance

The second part of this experiment aims at assessing how the DSMs perform the task of filtering out documents with a low level of relatedness. To do that, we injected different sets of out-of-domain documents, randomly selected from the Europarl corpus to the original INTELITERM subcorpora. More precisely, we injected 5%, 10%, 15% and 20%[9] to the various subcorpora. As we can see in Figure 2, the more noisy documents are injected, the lower is the NCT. Then, the methodology described in Section 4 was applied to these "new twelve subcorpora" (int_en05, int_en10, ..., int_it15 and int_it20, see Figure 2). As a result, at this point we have the documents ranked in a descending order according to their DSMs scores.

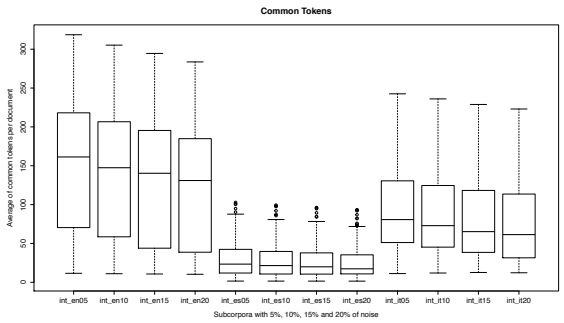

Figure 2: Average scores between documents when injecting 5%, 10%, 15% and 20% of noise to the various subcorpora.

In order to evaluate the DSMs precision, we analysed the first n positions in the ranking lists produced by the three DSMs (individually), and in this case n is the number of original documents in a given INTELITERM subcorpus. Table 3 presents the precision values obtained by the DSMs when injecting different amounts of noise to the various original subcorpora.

SubC	Noise	NCT	SCC	χ^2
int_en	*5%*	0.89	0.22	**1.00**
	10%	0.73	0.33	**1.00**
	15%	0.73	0.36	0.95
	20%	**0.80**	0.37	**0.90**
int_es	*5%*	0.00	0.00	0.38
	10%	0.07	0.07	0.20
	15%	0.09	0.09	0.17
	20%	0.14	0.18	0.23
int_it	*5%*	0.88	0.13	0.88
	10%	0.82	0.06	0.82
	15%	0.74	0.09	0.83
	20%	**0.73**	0.13	0.87

Table 3: DSMs precision when injecting different amounts of noise to the various subcorpora.

As expected, none of the DSMs got acceptable results for Spanish, being incapable of correctly identify noisy documents. However, we need to be aware that this happened due to the pre-existing low level of relatedness between the original documents in the int_es subcorpus (see Section 5.1 for more details). On the other hand, the DSMs show promising results for English and Italian. By

[9]The number of documents that correspond to these percentages can be inferred from Table 1.

a way of example, the χ^2 was capable of reaching 100% when injected 5% and 10% of noise to the int_en subcorpus, and even 90% when injected 20%. Although the NCT got lower precision, in general, when compared with the χ^2, it still reached 80% and 73% when injected 20% of noise to the English and to the Italian subcopora, respectively. From the evidences shown in Table 3, we can say that the NCT and the χ^2 are suitable for the task of filtering out low related documents with a high precision degree. The same cannot be say to the SCC measure, at least for this specific task.

6 Conclusions and Future Work

In this paper we presented a simple methodology and studied various Distributional Similarity Measures (DSMs) for the purpose of measuring the relatedness between documents in specialised comparable corpora. As input for these DSMs, we used three different input features (lists of common tokens, lemmas and stems). In the end, we conclude that for the data in hand these features had similar performance. 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 it is possible to assess and describe comparable corpora through statistical methods. The number of entities shared by their documents, the average scores obtained with the SCC and the χ^2 measure resulted to be an important surgical toolbox to dissect and microscopically analyse comparable corpora.

Furthermore, these DSMs can be seen as a suitable tool to rank documents by their similarities. A handy feature to those who manually or semi-automatically compile corpora mined from the Internet and want to retrieve the most similar ones and filter out documents with a low level of relatedness. Our findings show promising results when filtering out noisy documents. Indeed, two of the measures got very high precision results, even when dealing with 20% of noise.

In the future, we intend not only to perform more experiments with these DSMs in other corpora and languages, but also test other DSMs, like Jaccard or Cosine and compare their performance.

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 nº 317471. The research reported in this work has also been partially carried out in the framework of the Educational Innovation Project TRADICOR (PIE 13-054, 2014-2015); the R&D project INTELITERM (ref. nº FFI2012-38881, 2012-2015); the R&D Project for Excellence TERMITUR (ref. nº HUM2754, 2014-2017); and the LATEST project (ref. 327197-FP7-PEOPLE-2012-IEF).

References

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

Douglas Biber. 1988. *Variation across speech and writing*. Cambridge University Press, Cambridge, UK.

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.

Gloria Corpas Pastor. 2001. Compilación de un corpus ad hoc para la enseñanza de la traducción inversa especializada. *TRANS, Revista de Traductología*, 5(1):155–184.

Hernani Costa, Hugo Gonçalo Oliveira, and Paulo Gomes. 2010. The Impact of Distributional Metrics in the Quality of Relational Triples. In *19^{th} 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 *15^{th} Portuguese Conf. on Artificial Intelligence*, volume 7026 of *EPIA'11*, pages 597–609, Lisbon, Portugal, October. Springer.

Hernani Costa, Hanna Béchara, Shiva Taslimipoor, Rohit Gupta, Constantin Orasan, Gloria Corpas Pastor, and Ruslan Mitkov. 2015.

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. 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.

Hernani Costa. 2015. Assessing Comparable Corpora through Distributional Similarity Measures. In *EXPERT Scientific and Technological Workshop*, pages 23–32, Malaga, Spain, June.

EAGLES. 1996. Preliminary Recommendations on Corpus Typology. Technical report, EAGLES Document EAG-TCWG-CTYP/P., May. `http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html`.

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.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *MT Summit*.

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.