Cross-Language High Similarity Search Using a Conceptual Thesaurus

Parth Gupta, Alberto Barrón-Cedeño, and Paolo Rosso

Natural Language Engineering Lab. - ELiRF
Department of Information Systems and Computation
Universitat Politècnica de València, Spain
{pgupta,lbarron,prosso}@dsic.upv.es
http://www.dsic.upv.es/grupos/nle

Abstract. This work addresses the issue of cross-language high similarity and near-duplicates search, where, for the given document, a highly similar one is to be identified from a large cross-language collection of documents. We propose a concept-based similarity model for the problem which is very light in computation and memory. We evaluate the model on three corpora of different nature and two language pairs English-German and English-Spanish using the Eurovoc conceptual thesaurus. Our model is compared with two state-of-the-art models and we find, though the proposed model is very generic, it produces competitive results and is significantly stable and consistent across the corpora.

1 Introduction

The task of high similarity search refers to the identification of documents that are duplicates or share almost identical information. The proliferation of information in the age of the Web is extremely high and there exists a large redundancy in the contents of newly generated text. High similarity search becomes important either to avoid or to exploit redundancy. The former refers to the technology of duplicate identification for Web search indexing, also known as near-duplicate detection; whereas the latter corresponds to high similarity search for text classification, document clustering, plagiarism detection and retrieval by example. This problem is well studied for the monolingual variant and the most popular approaches are related to shingling [1], and the majority of research is based on the selection of a representative signature for the documents in question [2,3,4].

Documents with similar content also exist across languages, e.g. Wikipedia articles in multiple languages, news stories in different languages covering the same event, cross-language cases of plagiarism, and translated documents. Identification of such documents across languages is also referred as cross-language (CL) high similarity search, CL near-duplicate identification and CL pairwise similarity in the literature, but has attained less attention compared to its monolingual counterpart [5,6].

Usually, in this framework the length of the query is quite large (i.e. a whole document). Although it induces more information for the similarity estimation, this may potentially introduce noise. Moreover, the CL setting, where one term in language L_1

T. Catarci et al. (Eds.): CLEF 2012, LNCS 7488, pp. 67–75, 2012.

may stand equivalent to many completely different terms in language L_2, in addition to a large reference collection, introduces a new twist in the problem. The large vocabulary of the collection is *dangerous* in terms of ambiguity and computational cost.

We propose an algorithm which measures the CL similarity based on a conceptual thesaurus (CT). The main contributions of this work are twofold:

1. A method to represent documents (of any domain) in the conceptual space using a domain specific CT is suggested.
2. A novel method for CL high similarity search based on a reduced vocabulary (concepts) is proposed.

The rest of the paper is structured as follows. In Section 2, we present the related work. Section 3 describes the CT used and the models in detail. In Section 4, we present the performance evaluation of the approach and analysis. Finally, in Section 5 we summarise the work.

2 Related Work

Recently, there have been many attempts to address the issue of CL high similarity search. Anderka et al. [5] discuss the fact that, linear scan is inevitable for CL high similarity search, empirically and theoretically but do not report experimental results of the actual retrieval. Ture et al. [6] report the results of locality-sensitive hashing scheme [1] for the specified problem using MapReduce [7] and conclude as no optimal solution to reduce the search space. Moreover, they concentrate more on the scalability issues of the problem. In another approach, Platt et al. [8] suggest an oriented principle component analysis (OPCA) based learning in which multilingual documents are represented in a common space, but as they further mention, this technique is impractical for large vocabularies because the temporal and spatial cost scale quadratically with the vocabulary size.

Eurovoc has previously been used for the identification of translated documents [9,10], in which, the Eurovoc concepts were enriched by a set of associative phrases extracted from a large manually (keywords) annotated corpus. The Eurovoc concepts are then assigned to the documents based on the similarity between the contents of the document and the enriched associative set. This approach is quite restrictive because it demands a large manually annotated and domain dependent corpora for the association of Eurovoc concepts to the documents.

The CL explicit semantic analysis model (CL-ESA) tries to estimate the semantic similarity between two documents based on a comparable corpus [11]. The CL alignment based similarity analysis model (CL-ASA) is an adaptation of IBM M1 [12], in which the translation model is adapted to handle long texts and the language model is substituted by a length model [9] to measure the similarity [13,14]. Another model is based on the comparison of character n-grams (CL-CNG) between the documents [15]. Recently, these three models were compared in [16]. CL-ASA and CNG showed better performance on different corpora like JRC and Wikipedia. Therefore, we compare the proposed model with CL-ASA and CL-CNG.

3 Models

In this Section we describe our proposed model as well as the models we compare it with. The proposed model tries to measure the similarity between the documents in terms of shared concepts, assigned using a CT, and named entities (NEs) among them.

3.1 Conceptual Thesaurus

A CT contains concepts that are often multi-word structures and exhaustively try to cover the omnipresent concepts of the specific domain. The CT we use is Eurovoc[1], which has emerged from European Parliamentary proceedings. Eurovoc is a thriving resource and contains 6,797 multilingual concepts maintained with comparable identifiers (*concept id*) in 22 languages, which span across 21 domains of European Parliament activities. Some of the entries of Eurovoc are presented in Table 1.

Table 1. Examples of Eurovoc descriptors in the three languages

English	Spanish	German
action for failure to fulfil an obligation	recurso por incumplimiento	Klage wegen Vertragsverletzung
extra-community trade	intercambio extracomunitario	außergemeinschaftlicher Handel
sexual harassment	acoso sexual	sexuelle Belästigung

3.2 Cross-Language Conceptual Thesaurus Based Similarity (CL-CTS)

We represent the documents as a vector of the concepts in the thesaurus, rather than the original terms of the document. Concept assignment is the least trivial part. The concept assignment based on its verbatim occurrence in the document produces poor results [17]. Therefore, we assign a concept to a document if it "triggers the concept". Triggering is explained by the function $v(e, d)$ where e and d are Eurovoc concept and reference document respectively:

$$v(e, d) = \sum_{t \in e, T_e} f(t, d) \tag{1}$$

where, $f(t, d)$ depicts the frequency of term t in d. $\forall t \in e \cup d$ is stemmed and not a stopword. T_e refers to the vocabulary of Eurovoc concepts. The concept e is assigned to d with weight $v(e, d)$ if $v(e, d) > 0$.

We try to exploit this multilingual structure based on a heuristic: *the terms together are highly domain dependent but alone are domain independent*, e.g. "community" and "trade" may individually well be present in any domain compared to the complete descriptor "community trade". Moreover, we believe not all the terms help in the similarity estimation. Fig. 1 depicts the document frequency (df) of Eurovoc concept terms T_e,

[1] `http://eurovoc.europa.eu/`

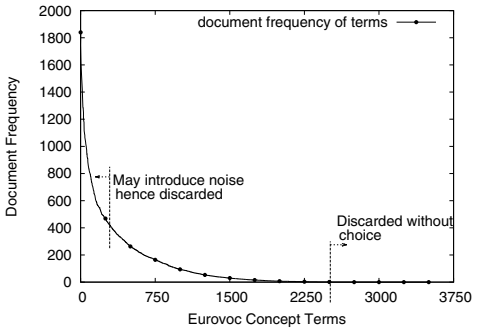

Fig. 1. Document frequency (in decreasing order) of the Eurovoc concept terms in the PAN subcorpus (cf. Section 4.1)

which is well in accordance with the Zipf's law. $\forall t \in T_e$, $df(t) = 0$ specifies that t does not participate in the similarity estimation. On the other hand, there are few terms for which, $df(t)$ is very high. These terms are less discriminative and, more importantly, very likely to introduce noise by increasing the similarity of non-relevant documents, especially when we use a reduced vocabulary. Therefore, we choose $t \in T_e$ for which $0 < df(t) < \beta$ as T'_e which is referred as reduced concepts (RC). In case of RC, T_e is replaced by T'_e in Eq. 1.

The conceptual vectors representing the documents are constructed on a monolingual basis, where each dimension represents one *concept id*. To find the similar documents for a given document q in language L_1, from the collection of documents D in language L_2, similarity between the conceptual vectors of q and $\forall d \in D$ is calculated as in Eq. 2, where $\boldsymbol{c}$ corresponds to the conceptual vector and $|\cdot|$ denotes *cardinality*.

$$\omega(q,d) = \frac{\alpha}{2} * \left(\frac{\boldsymbol{c_q} \cdot \boldsymbol{c_d}}{|q||d|} + \ell(q,d) \right) + (1-\alpha) * \zeta(q,d) \tag{2}$$

The first term is the conceptual component and the second is the named entity (NE) component. Here, $\zeta(\cdot,\cdot)$ defines the cosine similarity of char 3-grams between the NEs, $\ell(\cdot,\cdot)$ is the length factor (LF) penalty for the document pair as defined in [9] and α is the weighing factor so that $\omega(q,d) \in [0,1]$. The motivation behind the NE component is, NEs act as the discriminative features for the identification of different documents on the similar conceptual topics. To handle the variation of NEs across languages, we use character n-gram based similarity estimation. Moreover, the parallel documents follow a specific length distributions as specified in [9] that helps in incorporating the length information of parallel document pairs. Inclusion of LF induces this information in the similarity estimation.

3.3 Cross-Language Alignment Based Similarity Analysis (CL-ASA)

CL-ASA measures the similarity between two documents from different languages by estimating the likelihood of one document being a translation of the other one [13,14]. The similarity between the documents q and $d \in D$ is computed as in Eq. 3.

$$\omega(q, d) = \ell(q, d) * t(q \mid d) \tag{3}$$

where, $\ell(q, d)$ is again the length factor defined in [9] and the translation model $t(q \mid d)$ is calculated as in Eq. 4.

$$t(q \mid d) = \sum_{x \in q} \sum_{y \in d} p(x, y) \tag{4}$$

where, $p(x, y)$ is computed on the basis of a statistical bilingual dictionary which can be obtained from a parallel corpus.

3.4 Cross-Language Character n-Grams (CL-CNG)

The character n-grams have shown to improve the performance of cross-language information retrieval immensely for syntactically similar languages [15]. The documents are codified into the space of character n-grams and represented as the vectors of them. The CL-CNG measures the $\omega(q, d)$ as shown in Eq. 5.

$$\omega(q, d) = \frac{\boldsymbol{q'} \cdot \boldsymbol{d'}}{|q'||d'|} \tag{5}$$

where $\boldsymbol{q'}$ and $\boldsymbol{d'}$ are the projected vectors of q and d into character n-grams space.

4 Experiments and Analysis

We consider the documents in English as query documents $q \in Q$ and the documents in German or Spanish as reference documents $d \in D$. The aim is to find the highly similar document d for each q from D for each source language. In our experimental set up, there exists a highly similar document $d \in D$ for each q and the performance of the algorithms is evaluated in terms of the retrieval quality. We carry out the evaluation of the algorithms on three different datasets (Section 4.1) and two language pairs: English-Spanish (*en-es*) and English-German (*en-de*). We compare the proposed model with two state-of-the-art models, CL-ASA and CL-CNG. The results and analysis are presented in Sections 4.2 and 4.3 respectively.

4.1 Datasets

We extracted a collection of parallel documents from the JRC-Acqis corpus[2] referred as JRC, CL plagiarism cases from the PAN-PC-11 corpus[3] referred as PAN and Wikipedia comparable articles referred as Wiki for both language pairs. The JRC sub-corpus amounts to 10,000 documents for each language, PAN sub-corpus contains 2920 *en-es* and 2222 *en-de* document pairs and Wiki sub-corpus contains 10,000 documents for each language. The partitions of the JRC-Acquis and Wikipedia sub-collections

[2] http://optima.jrc.it/Acquis/

[3] http://www.uni-weimar.de/cms/medien/webis/research/corpora/pan-pc-11.html

used in the experiments are publicly available[4]. Our complete test collection includes 70,282 documents. The JRC corpus contains documents related to European Commission activities, while the PAN sub-corpus contains documents from Project Gutenberg[5]. Therefore, the vocabulary shared by Eurovoc and JRC is higher than that of Eurovoc and PAN or Wiki.

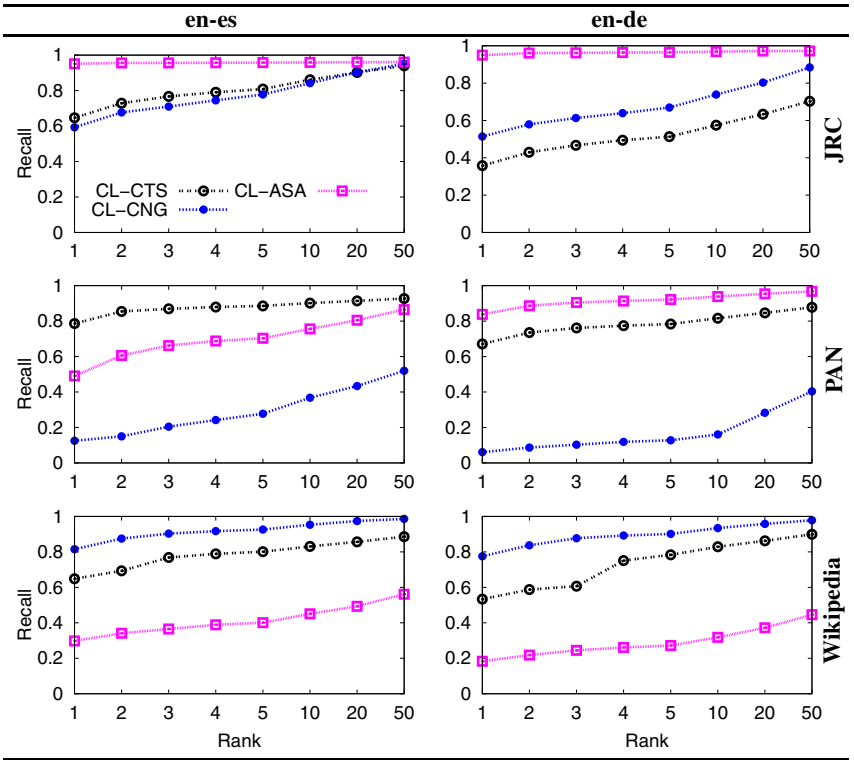

Fig. 2. Results of the proposed CL-CTS model on the JRC, PAN and Wiki sub-corpora and comparison with CL-CNG and CL-ASA. The performace is evaluated as Recall-over-Rank, where Recall@1 refers to the identification of the highly similar document at the very first position in the ranklist.

4.2 Results

We trained the translation model of CL-ASA on a different partition of JRC corpus of 10,000 parallel documents for each language pair and length factor values are used as suggested in [16]. The diacritics of Spanish and German are normalised for the similarity estimation in case of CL-CNG and n=3 is used. The parameters of CL-CTS, α and β, were set empirically on a small validation set of 500 documents from each corpus. We used $\beta = 0.10 * |D|$ for the three corpora while the $\alpha = 0.95$ for the JRC and α=0.50

[4] `http://users.dsic.upv.es/grupos/nle/downloads.html`
[5] `http://www.gutenberg.org/`

for the PAN and Wiki. Moreover, the LF is disabled on Wiki sub-corpus as the documents are not parallel. The performance of the models is measured by recall-over-rank as depicted in Fig. 2.

4.3 Analysis

The performance of CL-CTS with reduced concepts is much higher compared to the inclusion of all concepts because including the very common concepts increases the similarity score of some irrelevant documents. Let T'_{e,L_1} and T'_{e,L_2} denote the Eurovoc reduced concepts for language L_1 and L_2 respectively. The performance with RC heuristic will be driven by the size of $|T'_{e,L_1} \cap T'_{e,L_2}|$, which is usually quite high for parallel and comparable corpora, where $|\cdot \cap \cdot| = 1$ if both sets contain equivalent CT concepts in the respective languages. In general, the incorporation of NE component improves the performance except for JRC, which is very biased towards a particular category of NE as discussed later in this section. But this effect was minimised by the value of α = 0.95 for JRC. To handle the terms compounding in German we used jWordSplitter[6] which employs a greedy approach for splitting. Usually, German document retrieval stays more difficult compared to Spanish document retrieval for the word-based approaches because of the terms compounding.

Table 2. Average distribution of NEs in the three corpora

Corpus	Person	Location	Organisation	Total
JRC	1.8%	2.3%	8.7%	12.9%
PAN	1.8%	1.7%	1.9%	5.4%
Wiki	4.7%	3.7%	5.5%	14.0%

The other systems, CL-CNG and CL-ASA show very corpus dependent performance. To better describe this behaviour, we present the nature of these corpora and some statistics of the named entities[7] in the corpora in Table 2. JRC contains *parallel* documents of the European Commission activities which are highly domain dependent and contain quite large amount of NEs of type Organisation and Location (country names). These names appear quite identically in several documents. PAN contains cross-language plagiarism cases, which can be treated as *noisy parallel* data. These documents were generated using the machine translation technologies to translate text fragments from Project Gutenberg documents [16]. PAN documents are about literature and contain far more natural language terms compared to NEs. On the other hand, Wikipedia articles are *comparable* documents with a high amount of NEs. The amount and type of NEs in PAN and Wiki are quite diverse and balanced compared to JRC.

CL-ASA performs better on the JRC sub-corpus and very poor on the Wiki sub-corpus while CL-CNG performs better on the Wiki sub-corpus and very poor on the PAN sub-corpus. CL-ASA produces better results on nearly parallel data while CL-CNG demonstrates better performance on the NE dominated corpora. CL-CTS exhibits

[6] `http://www.danielnaber.de/jwordsplitter/`

[7] LingPipe NE Recogniser is used for English and Spanish; while, Stanford NER for German.

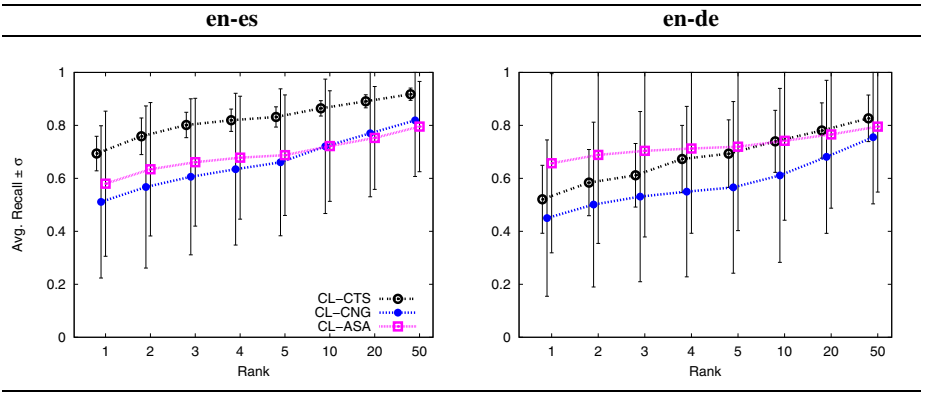

Fig. 3. Mean and standard deviation of the performance of the algorithms over different corpora

very stable performance across the corpora. The average performance of all the systems with their standard deviation is shown in Fig. 3. It is noticeable from the standard deviation values that CL-CTS is the most consistent across the corpora. CL-CTS can be very useful in the situation when the nature of the data is unknown or when dealing with a heterogeneous data. Moreover, CL-CTS uses a reduced vocabulary equals to $|T'_e|$ and NEs to measure the similarity between q and d. Other terms are discarded, resulting in very compact inverted index and a low computational cost. This reduces the temporal and spatial cost of the model dramatically. It should also be noted that CL-CTS achieves a stable performance across the domain with a domain specific conceptual thesaurus.

5 Summary and Future Work

We have proposed a model based on conceptual similarity for cross-language high similarity search which has very low temporal and spatial cost. The proposed model outperforms the character n-gram similarity based model on the linguistic sub-corpus PAN. The model also outperforms the machine translation based model on the comparable Wikipedia sub-corpus. The model demonstrates a very high stability across the corpora and performs consistently.

In future, we plan to test this model on a wide variety of language pairs, such as English with Hindi, Greek and Arabic. We also plan to compare the performance of this model to statistical conceptual models such as latent semantic analysis.

Acknowledgment. This work was done in the framework of the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems and it has been partially funded by the European Commission as part of the WIQ-EI IRSES project (grant no. 269180) within the FP 7 Marie Curie People Framework, and by the Text-Enterprise 2.0 research project (TIN2009-13391-C04-03). The research work of the second author is supported by the CONACyT 192021/302009 grant.

References

1. Broder, A.Z.: Identifying and Filtering Near-Duplicate Documents. In: Giancarlo, R., Sankoff, D. (eds.) CPM 2000. LNCS, vol. 1848, pp. 1–10. Springer, Heidelberg (2000)
2. Chowdhury, A., Frieder, O., Grossman, D., McCabe, M.C.: Collection Statistics for Fast Duplicate Document Detection. ACM Trans. Inf. Syst. 20, 171–191 (2002)
3. Charikar, M.S.: Similarity Estimation Techniques from Rounding Algorithms. In: Proceedings of the Thiry-fourth Annual ACM Symposium on Theory of Computing, STOC 2002, pp. 380–388. ACM, New York (2002)
4. Kolcz, A., Chowdhury, A., Alspector, J.: Improved Robustness of Signature-based Near-Replica Detection via Lexicon Randomization. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2004, pp. 605–610 (2004)
5. Anderka, M., Stein, B., Potthast, M.: Cross-Language High Similarity Search: Why No Sub-linear Time Bound Can Be Expected. In: Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Rüger, S., van Rijsbergen, K. (eds.) ECIR 2010. LNCS, vol. 5993, pp. 640–644. Springer, Heidelberg (2010)
6. Ture, F., Elsayed, T., Lin, J.J.: No Free Lunch: Brute Force vs. Locality-Sensitive Hashing for Cross-Lingual Pairwise Similarity. In: Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, pp. 943–952 (2011)
7. Dean, J., Ghemawat, S.: MapReduce: Simplified Data Processing on Large Clusters. Commun. ACM 51(1), 107–113 (2008)
8. Platt, J., Toutanova, K., tau Yih, W.: Translingual Document Representations from Discriminative Projections. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. EMNLP 2010, pp. 251–261 (2010)
9. Pouliquen, B., Steinberger, R., Ignat, C.: Automatic Linking of Similar Texts Across Languages. In: Recent Advances in Natural Language Processing III. Selected Papers from RANLP 2003, pp. 307–316 (2003)
10. Steinberger, R., Pouliquen, B., Hagman, J.: Cross-Lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC. In: Gelbukh, A. (ed.) CICLing 2002. LNCS, vol. 2276, pp. 415–424. Springer, Heidelberg (2002)
11. Potthast, M., Stein, B., Anderka, M.: A Wikipedia-Based Multilingual Retrieval Model. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 522–530. Springer, Heidelberg (2008)
12. Brown, P.F., Pietra, V.J.D., Pietra, S.A.D., Mercer, R.L.: The Mathematics of Statistical Machine Translation: Parameter Estimation. Comput. Linguist. 19, 263–311 (1993)
13. Barrón-Cedeño, A., Rosso, P., Pinto, D., Juan, A.: On Cross-lingual Plagiarism Analysis using a Statistical Model. In: Proceedings of the ECAI 2008 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse, PAN 2008 (2008)
14. Pinto, D., Civera, J., Barrón-Cedeño, A., Juan, A., Rosso, P.: A Statistical Approach to Crosslingual Natural Language Tasks. J. Algorithms 64, 51–60 (2009)
15. Mcnamee, P., Mayfield, J.: Character N-Gram Tokenization for European Language Text Retrieval. Inf. Retr. 7(1-2), 73–97 (2004)
16. Potthast, M., Barrón-Cedeño, A., Stein, B., Rosso, P.: Cross-Language Plagiarism Detection. Language Resources and Evaluation, Special Issue on Plagiarism and Authorship Analysis 45(1) (2011)
17. Pouliquen, B., Steinberger, R., Ignat, C.: Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus. CoRR **abs/cs/0609059** (2006)