Applying a Lexical Similarity Measure to Compare Portuguese Term Collections

Marcirio Silveira Chaves and Vera Lúcia Strube de Lima

Pontifícia Universidade Católica do Rio Grande do Sul - PUCRS
Faculdade de Informática - FACIN
Programa de Pós-Graduação em Ciência da Computação - PPGCC
Av. Ipiranga, 6681 - Partenon - Porto Alegre - RS
CEP 90619-900
{mchaves, vera}@inf.pucrs.br

Abstract. The number of ontologies publicly available and accessible through the web has increased in the last years, so that the task of finding similar terms[1] among these structures becomes mandatory. We depict the application and the evaluation of a new similarity measure for comparing Portuguese Ontological Structures (OSs) called Lexical Similarity (LS). This paper describes contributions to the study and application of mapping between terms present in multidomain OSs. In order to approach this mapping we combine preliminar similarity measures and heuristics. Our measure uses a stemmer, it is established upon String Matching (SM) proposed in [1] and it was evaluated by means of a comparison to human evaluation. Finally, we concentrate on the application of LS measure to terms belonging to same domain thesauri and discuss the results obtained.

Key words: Lexical Similarity Measure, Mapping, Ontological Structures

1 Introduction

The automatic mapping between Ontological Structures (OSs) has been a continuous concern as a task of integration and reuse of knowledge. However, the manual execution of such task is quite tedious and slow, so it is important to automate it, at least partially.

In this work, OSs are understood as sets of pre-defined terms explicitly connected by semantic relations in a format, which is readable by humans and machines. This notion is suitable for collections of vocabularies as well as for collections of concepts.

Several efforts have been reported in the literature to mapping different OSs in English language [2,3,4] and in German language [1]. However, other works

[1] The words “terms” and “concepts” will be used with the same meaning in this article.

that deal with Portuguese OSs have not been found. We concentrate our efforts on Portuguese OSs, developing, testing, validating and evaluating a proper measure to help detecting similar terms between OSs, which are projected independently using preview studies [1,3].

This paper is further organized as follows. Section 2 describes the SM measure [1]. Section 3 details the similarity measure proposed in this paper. The experiments accomplished over multidomain Portuguese OSs are presented in Section 4. Section 5 presents the experiments with thesauri belonging to the same domain. Finally, Section 6 gives an outlook on future work.

2 Maedche and Staab Measure

Maedche and Staab [1] present a two layer approach, first lexical and then conceptual, to measure the similarity between terms of different OSs. At the lexical level, they consider the Edit Distance (ED) formulated by Levenshtein [5]. This distance contemplates the minimum number of insertions, deletions or substitutions (reversals) necessary to transform one string into another using a dynamic programming algorithm. The contribution of Maedche and Staab consists of the String Matching (SM) measure given by:

$$SM(T_i, T_j) := max\left(0, \frac{min(\mid T_i \mid, \mid T_j \mid) \; - \; ED(T_i, T_j)}{min(\mid T_i \mid, \mid T_j \mid)}\right) \; \in \; [0, 1] \; . \qquad (1)$$

The SM measure calculates the similarity between two terms (Ti, Tj). The length in characters of the shortest term is represented by $min(|Ti|, |Tj|)$. For example, to obtain the similarity between the terms (`comerciario`, `comerciante`) the minimum length is 11 and $ED(Ti, Tj)$ is 3 (changes “r” by “n” and inserts “t” and “e”). Thus, the resulting value for SM(`comerciario`, `comerciante`) is 0.73.

This measure always returns a value between 0 and 1, where 1 stands for perfect match and zero indicates absence of match. Maedche and Staab worked with German language OSs from tourism domain. However, while applying SM measure to Portuguese OSs, many terms were mapped inconsistently. In order to get better results we developed a proper measure, which was validated and evaluated[2].

3 Lexical Similarity Measure

We propose an alternative to SM measure which is based on the radicals[3] of the words. Generally, these radicals are the most representative part of a word in Portuguese, and they can be extracted with the help of a stemmer. We used a

[2] Detailed results, experiments, validation and evaluation can be found in [6].

[3] The term radical as used in this article represents the initial character string of a word and not necessarily the linguistic concept of radical.

stemmer that was specifically developed for Portuguese by Orengo and Huyck, which presented good performance when compared [7] to Porter algorithm or other [8]. Our proposal is named Lexical Similarity (LS) and it is expressed by the equation in 2, where terms are represented by T_i and T_j, and index i points to the terms in OS$_A$ while index j refers to terms in OS$_B$.

$$LS(T_i, T_j) = min\{\mathit{\Delta}^1_{ij}, \mathit{\Delta}^2_{ij},\ \ldots, \mathit{\Delta}^k_{ij}\}\ \in\ [0,1]\ . \tag{2}$$

Terms can be formed by single-words, or by more than one word. LS measure, in contrast to SM measure, considers only the radical of each word, instead of the complete string of characters. The symbol $\mathit{\Delta}$ represents the value obtained by SM measure under the following conditions:

$$\mathit{\Delta}^k_{ij} = \begin{cases} SM(Rad^k_i, Rad^k_j) & if\ ED(Rad^k_i, Rad^k_j) = 0 \\ SM(Rad^k_i, Rad^k_j) - 0.1 & if\ ED(Rad^k_i, Rad^k_j) = 1 \\ SM(Rad^k_i, Rad^k_j) - 0.2 & if\ ED(Rad^k_i, Rad^k_j) = 2 \\ 0 & if\ ED(Rad^k_i, Rad^k_j) \geq 3 \end{cases} \tag{3}$$

The radical of a word that is part of a term T is represented by Rad^k_i, where k indicates the position of this word in T and i indicates the OS to which this term belongs. When T_i and T_j are multiword terms, the index k reaches the value of the amount of words of the term with the minimum number of words, so that LS measure calculates the similarity between the first k pairs of radicals (Rad^k_i, Rad^k_j) in the terms being compared.

The result returned by LS measure is the minimum value produced by equation 3, which depends on the Edit Distance. As the radical of a term owns a strong semantic weight, the result obtained by ED is decremented according to the conditions stated in equation 3. The highest is the ED, the highest is the penalty used. The penalty values (0.1 and 0.2) were obtained from empirical studies with SM measure. We assume that, if ED $\geq$ 3 the value returned by SM is zero and, consequently, LS is zero, too. What means, three or more changes in the radical of a word suggest a low degree of similarity.

For example, in order to check the similarity between the terms `areaEstrategica` and `armaEstrategica`, the words of the each term are processed by a stemming algorithm, which produces the stems “are” and “arm”, “estrateg” and “estrateg”, so that:

$$\begin{aligned} LS(areaEstrategica, armaEstrategica) = min\{&SM(are, arm), \\ &SM(estrateg, estrateg)\}\ . \end{aligned}$$

To calculate SM*(are, arm)*, we obtain the length of the shortest term, in this case 3. Then ED*(are, arm)* is calculated, which gives 1, since the letter “e” is changed to “m” to transform the string “are” into “arm”. So, SM*(are, arm)* is solved as:

$$max\left(0, \frac{3\ -\ 1}{3}\right)\ =\ 0.67\ .$$

As in this case $ED = 1$, the penalty to be applied is 0.1. So, the resultant similarity is 0.57.

The next result to be obtained is the similarity between SM(*estrateg, estrateg*) that is 1. In this case ED(*estrateg, estrateg*) is zero, (since the strings are in perfect match). Thus:

$$LS(areaEstrategica, armaEstrategica) = min\{0.57, 1\} = 0.57\ .$$

We did not find other works in the literature that provide a study on semantic weighting for each single-word in a multiword term, which would be suitable for Portuguese language as well as for several other languages such as Spanish, French and so on. In our proposal, as the reader can observe, words with the lowest lexical similarity value may perform an important role on similarity detection.

4 Multidomain Experiment

The OSs we used in this experiment come from two distinct sources[4]. Their terms belong to one of two groups: single-word terms or multiword terms[5].

The experiments were organized in two steps: testing and validation[6] of LS measure, followed by its evaluation. The terms in OS$_A$ were categorized into two sets for each phase, while terms in OS$_B$ remained without categorization during both validation and evaluation phases. The terms were placed in alphabetical order and an algorithm was developed to randomly distribute them through validation and evaluation experiment groups.

We also disclosed a heuristic to tune the mappings generated by LS measure. In Portuguese language, the semantic weight of the first characters in a term is apparently strong, which gives rise to the heuristic that is stated as:

$$If\ Rad[1]_i^k \neq Rad[1]_j^k\ then\ SM(Rad_i^k, Rad_j^k) = 0 \tag{4}$$

According to LS measure (equation 2), let the index inside the brackets be the position of the first character in the radical of the word in a term. If the two radicals Rad_i^k, Rad_j^k being compared have a different first letter, the value returned by SM measure will be zero. Consequently, LS will be zero, too.

[4] Namely: Brazilian Senate Thesaurus (OS$_A$) and São Paulo University - USP Thesaurus (OS$_B$).

[5] For the experiments with multiword terms, OSs were first preprocessed in order to eliminate blanks. Moreover, the first character of each word was capitalized, except for the first word in a term. This procedure is necessary to compare results with those in English [3] and German [1] experiments.

[6] Details on the experiments carried out in testing and validation can be found in [9].

For the evaluation phase, we used 1,823 single-word terms of Senate OS, while the USP OS remained with its original 7,039 single-word terms. We selected 4,701 multiword terms of Senate OS and kept 16,986 multiword terms of USP. The aim of the experiments in this phase was to check the agreement among LS and SM measures according to the results given by a human analysis of similarity.

In order to examine in detail the 2,887 pairs of terms and the corresponding system-computed or human confirmed analysis, we split them into seven groups. These groups are presented in Table 1, where G1 to G7 stand for the respective group[7].

Table 1. Composition of the groups according to a human point of view

	$SM \geq 0.75$ $LS \geq 0.75$	$SM \geq 0.75$ $LS < 0.75$	$SM < 0.75$ $LS \geq 0.75$
Terms estimated as similar by human analysis	G1	G2	G3
Terms estimated as unlike by human analysis	G4	G5	G6
Doubt	G7		

Human analysts pointed the pairs of terms as "similar", "unlike" or "doubtful". This result was compared with the automatically processed combinations. We choose Group G5 in Table 1 deemed as the most representative to be described in detail in the next section.

4.1 Analysis of Group G5

This group contains terms whose are deemed similar by SM measure and unlike by LS measure as well as by the human analysis. Moreover, in G5 there are most of the pairs analyzed during the evaluation phase, that is, about 73% which corresponds to 907 single-word terms and 1,211 multiword terms. We show an extract of these terms in Table 2.

Table 2 contemplates single-word (first five lines) and multiword (next five lines) terms. At first, let's analyze single-word terms. Most of them belonging to this group have the same suffix, that is, the final string is a perfect match of characters. As SM equally weights the strings belonging to the radical or to the suffix, a high value of similarity was observed between the terms having same suffix. However, this policy is not yet confirmed for Portuguese.

Otherwise, in the multiword terms, at least one word of the term has the same suffix. As the reader may note, all terms in Table 2 seem to be unlike, despite SM measure detects them as similar. We can increase the threshold from 0.75 to 0.8 in order to get a more consistent mapping by SM. However, this higher threshold is not enough to deem the terms belonging to G5 as dissimilar, once just some pairs of terms have similarity value under 0.8.

[7] We used the threshold 0.75 in our experiments. This value is also used in [1].

Table 2. Extract of group G5: single-word and multiword terms

OS_A	OS_B	SM	LS
tuberculose	tuberculos	0.90	0.5
terceiros	terreiros	0.89	0.65
atentado	atestado	0.88	0.70
corretor	corredor	0.88	0.65
desarmamento	desmatamento	0.75	0
delitoFiscal	debitoFiscal	0.92	0.70
ensinoMedico	ensinoMedio	0.91	0.65
policiaAdministrativa	politicaAdministrativa	0.90	0.65
direitoPenalEcologico	direitoPenalEconomico	0.90	0.47
direitoAVida	direitoAVoto	0.75	0.13

As this group represents most of the terms analyzed in evaluation phase and, taking into account the results generated by SM measure, it is possible to question if this measure is really proper to treat Portuguese terms. Specifically for multiword terms, we believe that the best performance of LS measure is due to the fact that it considers each constituent word individually.

As a following step toward experimentation, we concentrate our efforts in mapping of terms belonging to the same domain. We apply the SM and LS measures to these terms through the experiment described in the next section.

5 Same Domain Experiment

In this experiment we verify the similarity among 2,083 terms from GEODESC Thesaurus[8] and 429 terms from USP Thesaurus, which belong to the Geosciences domain. In order to carry out this experiment, we do not consider the cases where there is a perfect matching of characters, because these ones do not help to evaluate any of the measures. Moreover, we use the first letter heuristic to help us obtain better results.

After running the algorithm with the two measures, 91 mappings were found between the two thesauri representing 4.36% of the terms of GEODESC Thesaurus and 21.21% of the terms of USP Thesaurus. In order to analyze these mappings, we split them into 2 groups. In Group A (GA) these are the terms considered similar by LS measure, while the Group B (GB) includes the terms deemed as similar by SM and dissimilar by LS. Table 3 shows these groups considering similar terms with similarity value ≥ 0.75.

Table 3 presents the combinations between SM and LS similarity measures. These cases are explained as follows:

[8] Available by ftp://ftp.cprm.gov.br/pub/pdf/didote/geodesc.pdf

Table 3. Groups composed after same domain experiment

Group	Conditions	
GA	SM < 0.75	LS ≥ 0.75
	SM ≥ 0.75	LS ≥ 0.75
GB	SM ≥ 0.75	LS < 0.75

5.1 Analysis of Group A

This group contains those terms which are considered similar by LS measure. The analysis was broken into two tables, comparing our LS measure with Maedche and Staab's SM measure. Only 4 mappings were detected while considering SM < 0.75 and LS ≥ 0.75, as is shown in Table 4. In our point of view just the first mapping (between the terms `sais` and `sal`) can be considered correct by LS. In order to evaluate the remaining mappings it is necessary to know the semantic relations among the terms and to take into account the meaning of each term.

Table 4. Pairs of terms considered dissimilar by SM and similar by LS

GEODESC	USP	SM	LS
sais	sal	0.33	1.00
arqueamento	arqueano	0.63	1.00
meteorito	meteoritica	0.67	0.76
vulcanicas	vulcanismo	0.70	1.00

In Group A, when both measures consider the terms being compared as similar (SM $\geq$ 0.75 and LS $\geq$ 0.75) we have the terms presented in Table 5. Lines 1 to 5 show terms with number variation and they are correctly deemed as similar by both measures. The remaining pairs of terms, such as those in Table 4, do not present a unique characteristic and it is difficult to perform an evaluation of the results generated.

5.2 Analysis of Group B

This group presents most of the mappings found in our experiment. We split these pairs of terms into two tables, the former composed by only one word terms and the latter by multiword terms.

The single-word terms are shown in Table 6. Despite all these pairs of terms have high lexical similarity, their meanings are different. So, in the context of mapping of similar terms between OSs we consider they should not be mapped.

In this moment it is important to stress a contribution of our measure. According to the literature studied, just the SM measure has been used to map terms among OSs. In this work, when we apply SM measure to single-word

Table 5. Pairs of terms considered similar by SM and LS

GEODESC	USP	SM	LS
lava	lavas	0.75	1.00
aguaSubterranea	aguasSubterraneas	0.87	1.00
depositosGlaciais	depositoGlacial	0.80	1.00
fumarola	fumarolas	0.88	1.00
oolitos	oolito	0.83	1.00
andesina	andesito	0.75	0.76
dolomita	dolomito	0.88	1.00
metamorficas	metamorfismo	0.75	1.00
metassomaticas	metassomatismo	0.79	0.79
prospeccaoGeotermal	prospeccaoGeotermica	0.84	1.00

Table 6. Single word pairs of terms

GEODESC	USP	SM	LS	GEODESC	USP
bioestratigrafia	litoestratigrafia	0.88	0.66	bioestratigraf	litoestratigraf
biologia	geologia	0.75	0.47	biolog	geolog
cosalita	sodalita	0.75	0.51	cosalit	sodalit
gemologia	geologia	0.88	0.73	gemolog	geolog
hamarita	hematita	0.75	0.51	hamarit	hematit
paleoecologia	paleontologia	0.85	0.62	paleoecolog	paleontolog
pedologia	geologia	0.75	0.47	pedolog	geolog
pinita	pirita	0.83	0.70	pinit	pirit
reologia	geologia	0.88	0.73	reolog	geolog
teleprocessamento	geoprocessamento	0.81	0	teleprocess	geoprocess

terms the reader can note its low performance, while our measure seems to attribute a suitable similarity value to the same pairs of terms. So, LS measure contributes to avoid detection of dissimilar terms like similar.

Still in this group, we analyze the multiword terms. The pairs of terms in this case are depicted in Table 7.

Table 7. Multiword pairs of terms

GEODESC	USP	SM	LS
faciesSedimentares	rochasSedimentares	0.78	0
geologiaAplicada	hidrologiaAplicada	0.75	0
geologiaEconomica	geologiaIsotopica	0.76	0
geologiaEstrutural	petrologiaEstrutural	0.83	0
geologiaFisica	geodesiaFisica	0.79	0
prospeccaoGeoquimica	prospeccaoBioquimica	0.90	0.51
sistemasOperacionais	sistemasDeposicionais	0.75	0

The reader may note that these pairs are considered similar by SM measure mainly due to the fact of dealing with them as a single string. As oppose to the LS measure, SM does not verify the similarity among individual words. The multiword terms belonging to Geosciences domain are generally composed by more than 10 characters. So, the value returned by ED does not generate sufficient impact to reduce the final similarity value of SM of the full term.

On the other hand, our measure considers individually the words belonging to the terms. This fact helps reducing the final similarity value, once the shortest term has a lower value than the one used by SM. So, the result of ED has a greater impact in the equation, decreasing the value of LS measure.

It is important to observe in Table 7 that most of the values generated by LS measure is zero. This occurs because those pairs have 3 or more distinct characters in the radical of the words.

Finally, it is worth noting the contribution of the penalties introduced in equation 2, as expressed in Table 8.

Table 8. Contribution of the penalties.

GEODESC	USP	SM	LS
bioestratigrafia	litoestratigrafia	0.88	0.66
lazurita	azurita	0.86	0.73
litoestratigrafia	bioestratigrafia	0.88	0.66
reologia	geologia	0.88	0.73

These penalties allow decreasing the value of LS measure and, consequently, considering terms as dissimilar (maintaining threshold 0.75), in opposite to SM measure. For example, the similarity between the pair of terms `bioestratigrafia` and `litoestratigrafia` by LS measure without penalties would be 0.86. This value allows us to consider it as similar, however, introducing the penalties (in this case 0.2) we have the final similarity value 0.66, which is under the threshold established. In fact, this pair is not really similar likewise the remaining ones in Table 8. Thus, they should not be mapped in the context of our analysis.

6 Final Remarks and Future Work

This work is the first effort towards the detection of similar terms between Portuguese OSs. LS measure was evaluated based on human evaluation of similarity, even though we find difficulties to evaluate similarity measures in agreement with a human point of view. A full description and analysis of the results obtained with LS measure are given in [6]. We believe that our measure contributes to help the ontology engineers reuse the information contained in the ontological structures, since the reuse is one of the main concerns in the context of the semantic web.

We carried out experiments with terms belonging to multidomain as well as to the same domain structures, and we commented the main results obtained. In spite of being them preliminary results, they are encouraging.

The next step is the application of LS measure to other languages, such as English or Spanish. In this situation a proper stemming algorithm, suitable for each different language, should be used. Besides, the similarity measures presented in this article can be used in order to aid on the task of union or alignment of ontological structures. It could also be connected to specific interface to help the ontologists detect terms suggested as similar.

Acknowledgements

Marcirio Silveira Chaves was supported by the research center HP-CPAD (Centro de Processamento de Alto Desempenho HP Brasil-PUCRS).

References

1. Alexander Maedche and Steffen Staab. Measuring Similarity between Ontologies. In *Proceedings of the European Conference on Knowledge Acquisition and Management - (EKAW-2002). Madrid, Spain, October 1-4*, pages 251–263, 2002.
2. AnHai Doan, Jayant Madhavan, Pedro Domingos, and Alon Halevy. Learning to Map between Ontologies on the Semantic Web. In *Proceedings of the World-Wide Web Conference (WWW-2002), Honolulu, Hawaii, USA*, May 2002.
3. Natalya Fridman Noy and Mark A. Musen. Anchor-PROMPT: Using Non-Local Context for Semantic Matching. In *Proceedings of the Workshop on Ontologies and Information Sharing at the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-2001), Seattle, WA*, August 2001.
4. Sushama Prasad, Yun Peng, and Timothy Finin. Using Explicit Information To Map Between Two Ontologies. In *Proceedings of the 1^{st} International Joint Conference on Autonomous Agents and Multi-Agent Systems - Workshop on Ontologies in Agent Systems (OAS) - Bologna, Italy. 15-19 July*, 2002.
5. Vladimir Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Cybernetics and Control Theory*, 10(8):707–710, 1966.
6. Marcirio Silveira Chaves. Comparação e Mapeamento de Similaridade entre Estruturas Ontológicas. Master's thesis, PUCRS-FACIN-PPGCC, 2004.
7. Viviane Moreira Orengo and Christian Huyck. A Stemming Algorithm for Portuguese Language. In *Proceedings of Eigth Symposium on String Processing and Information Retrieval (SPIRE-2001)*, pages 186–193, 2001.
8. Marcirio Silveira Chaves. Um Estudo e Apreciação sobre Dois Algoritmos de Stemming para a Língua Portuguesa. Jornadas Iberoamericanas de Informática. Cartagena de Indias - Colômbia (CD-ROM), August 11-15, 2003.
9. Marcirio Silveira Chaves and Vera Lúcia Strube de Lima. *Looking for Similarity between Portuguese Ontological Structures.* In: António Branco, Amália Mendes, Ricardo Ribeiro (editors). Edições Colibri, Lisboa, 2004 (to appear).