2009 International Conference on Computer Engineering and Applications
IPCSIT vol.2 (2011) © (2011) IACSIT Press, Singapore

Two-way Dictionary-based Lexical Ontology Alignment

Ahmad Adel Abu Shareha[+], Manadava Rajeswari, Dhanesh Ramachandram[1]

[1]Computer Vision Research Group, School of Computer Science, Universiti Sains Malaysia

Abstract. This paper introduces two-way dictionary based words/strings matching technique for ontology alignment. The proposed technique represents a step ahead into semantic ontology alignment. This technique uses a combination of approximate, exact, NLP method and error correction routines for ontology alignment. The overall design is based on the assumption that the strings in the ontology expose meanings. This meaning can be identified directly if the string has an equal corresponding dictionary entity or requiring some preprocessing if the string has no corresponding dictionary entity. For the non-dictionary entities, there are so many clarifications such as: misspelled word, compound word, foreigner language word, etc. These cases can be further clarified using the error correction technique which retrieves the most relevant words based on some approximate string matching and using lexical resource or a dictionary. All the extracted words then are Lemmatized and compared to each other in an exact string matching.

Keywords: Ontology alignment, lexical ontology alignment, semantic string matching.

1. Introduction

The ontology alignment problem is concerned with constructing a correspondence dialogue between two or more ontologies by discovering and matching their identical elements. [1] Matching the ontologies' elements is established either based on their similarities in so called schema-based or based on the similarities of some text instances that are provided for each element in so called instance-based approach. Instance based is achieved using lexical methods only [2, 3]. Schema based alignment is commonly used because of its capabilities to fit in different domains and applications because it requires no additional inputs and instances. The schema based alignment is achieved using two main approaches, the lexical approach and the structural approach. The structural approach for ontology alignment looks at the elements exchange relationships, affiliation and position in the structure. The lexical approach matches the elements based on their string properties (e.g.: names, labels). The lexical alignment which is of our interest is furthered classified into two main approaches, the syntactical and the semantic. In the syntactic approaches, string distance methods, both approximate and exact (e.g: edit-distance [4]) are used to calculate the similarity between the input strings. In the semantic, NLP methods are utilized which use the language rules and resources to semantically compare the input strings based on their semantics (meanings). Both, semantic and syntactic approaches are normally proceeded by some simple pre-processing like lower case conversion and string fragmentation based on some internal punctuation delimiters (e.g: space, upper case letter .. etc).

The syntactic-based alignment which depends on the approximate string distance methods is more flexible and widely used in the alignment systems. Several methods for string approximate distance have been developed for general purpose and then borrowed to be used in the alignment problem. These methods calculate the similarity/dissimilarity of the input strings' pairs and output a similarity value. Two problems are facing methods in the alignment application: First, the threshold value that works as a filtering mechanism that accepts or discards the pair as an identical pair based on their similarity value is critical and hard to be predetermined. Second, some absolutely different words have a very similarity string distance value, in such cases these methods fail in giving reasonable results even with a very sharp threshold. The

[+] Ahmad Adel Abu Shareha. Tel.: + (60173413734); fax: + (6046533888).
E-mail address: (adel@cs.usm.my).

semantic methods are very sensitive to the string representation; any changes that might occur, such as misspelling, will dramatically drop the performance of the semantic methods.

Thus, this paper introduces a technique for ontology alignment that combines and encapsulates the syntactic and semantic approaches and allows them to run simultaneously. Thus; we make use of the NLP method, approximate and exact matching based on dictionary. This technique does not require threshold, significantly discriminating the different words with close string representation and robust to words variation due to misspelling and compound words.

The rest of the paper is organized as follows: Section 2 discusses the related works both in methods and system prospective. The proposed technique is explained in Section 3. Section 4 shows the implementation details. The results discussion, significant achievements and setbacks are highlighted in section 5, suggestions based on the results are provided in section 6 which present the future work. Our conclusion is given in section 7.

2. Related Works

Several methods, approaches and systems have been developed for ontology alignment, most of these are based on the literature that exist in the schema matching, graph homomorphism, graph isomorphism, graph matching, taxonomy and others. In association to ontology alignment, the existing works can be categorized into methods and systems.

2.1. Methods

The schema-based, syntactic lexical ontology alignment depends solely on the string matching metrics. A clear example of such a method is the edit distance method [5] which calculates the dissimilarity of two input strings as the number of mutation steps that is required transforming one of the strings into the other. The mutation steps include insertion, deletion, substitution and transposition of characters with same or different costs for each mutation type. Other distances metrics have been developed such as Needleman-Wunch [6], Hamming distance [7], Winkler [8] which vary in the 'error model'. The 'error model' determines the type and the cost of the considered mutation steps. Apart from the general purpose string distance metrics mentioned earlier, Giorgos et al [9] have developed a metric for string distance especially for the ontology alignment. The proposed method combined some existing string distance methods that calculate both the similarity of the strings, the dissimilarity and used an existing string distance method, namely Winkler [8]. Overall, the syntactic methods (general or alignment specific) compare the strings blindly despite their semantics (meaning) which make them unable to cope with the word variations (e.g.: verb tenses). However, the ontology as knowledge representation is mostly human made which tends to have meaning for their string properties facing the fact the human tends to use and understand meaning than memorizing arbitrary strings.

The semantic approach is using language resources and rules to process strings that represent words and set up correspondence between these words [2]. The semantic approach is robust to word variations because it utilizes the language resources and rules to recognize and resolve such variations. Typical semantic methods are either intrinsic which use rules to implement semantic processing such as stemming with no outer resources like dictionary or extrinsic which uses lexical resources and dictionaries s. An example of the semantic methods that are used for alignment is the Synset matching. The synset matching uses lexical resource to discover the words synonym. Although, the semantic methods are robust to deal with words meanings and semantics, the semantic approach is unable to deal with unexpected slight variations of the words such as misspelling. This makes the semantic methods less efficient comparing to the syntactic methods for misspelling errors, compound words and other non-regular variation of the words which can be understood by the human and looks ambiguous for the machine.

2.2. Systems

In systems prospective, few lexical based systems have been developed which mostly commence by using string distance methods and end up with semantic methods or visa verse. In the hybrid ontology alignment systems (structural and lexical), lexical ontology matching has been used as a former step to determine the initial pairs that are used as a base for the structural alignment in later steps. In Cupid [10] the lexical alignment includes four steps for string processing, three of which are NLP methods: Normalization which includes tokenization based on punctuation, expansion and eliminations, categorization based on the elements data types (e.g: money, real .. etc) and finally apply the comparison based on the thesaurus and substring matching. LOM [11] use four methods for lexical matching those are: whole term exact matching, word constituent matching, synset matching using WordNet [12] and type matching using SUMO [13, 14] and MILO [15].

Previous work has concentrated on using the existing methods for string distance and NLP methods to build systems for ontology alignment; less effort has been given to develop methods before putting them in complete systems. Consequently, ontology in the alignment prospective has followed the schema and graph and the ontology string properties followed the records or database entries which are normally computer generated strings. In relation to methods, syntactic approaches cannot cope with the word variations and verb tenses, while semantic cannot cope with the misspelling errors and other unexpected errors. Thus, we propose to utilize a dictionary-based method with syntactic methods to solve such problems. The proposed solution utilizes both string distance and extrinsic semantic methods to set up correspondence between the ontologies' element. As a result the proposed solution utilizes the error-detection and correction mechanism that is used for the text, except that the inputs for such a mechanism in the proposed technique are single strings rather than with full text. In this work, the ontology has referred to its origin as knowledge representation which is human made. Humans tend to use semantic strings which might enclose some mistakes. Thus, the method developed here is designed to take care of such criteria.

3. Proposed Technique

In the Semantic Web, Artificial Intelligence and other fields, ontology is used as a knowledge source to resolve semantic problems. The applications that used ontology are clear proof that ontology is knowledge representation not only a simple data structure. Examples of such applications are: semantic indexing of the web in the semantic web and indexing and representation of multimedia contents. Thus, ontologies are representations of facts and logic [16]. Consequently, the alignment has to be a semantic mechanism where in some intelligent processes and rules have to be used.

So far, in the approximate string matching, the claim for efficiency has been stated for the methods and tools which can discriminate between two different words with a very similar shape such as 'winner' and 'winter' which is can clearly be resolved by using a dictionary or any lexical resources . But such a case is not the real problem that face the alignment, because the ontology are developed mostly by human and in which the developer used language words in different forms (e.g.: compound words, short cuts, etc….) to represent the labels of the ontology elements not an arbitrary strings which used for indexing of records purposes. Thus; the alignment needs an intelligent set of techniques that is able to deal with the ontology elements in an intelligent way. The proposed technique is a starting point which provides a single technique for ontology alignment which can be further used with a set of similar techniques to build an intelligent ontology alignment solution. The proposed technique takes the inputs as two groups of strings corresponding to two input ontologies. We assume that the strings in the ontologies are human made and intend to have meaning and might include some misspelling, mistakes, and compound words. The proposed technique is based on two well-known methods which are dictionary look up and error correction which intern use exact and approximate matching. Before starting the comparison process, first we check the input strings against the dictionary entries to trace the word units. This process is achieved using the dictionary look up technique. Based on this step, each group of strings is furthered classified into two groups: words whose components have corresponding entities in the dictionary and non-words whose components have no corresponding entities in the dictionary.

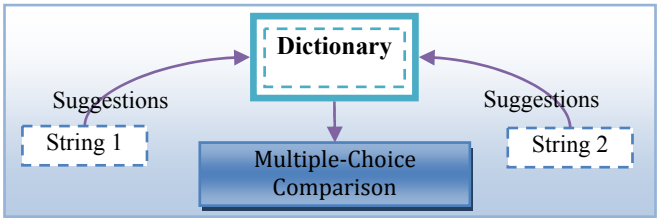

Fig. 1: Two-way comparison

For the *word components*, the comparison is a straight forward mechanism. Each word from the string group is compared to all the words in the other group that correspond to the other ontology. The variations of the words are handled using lemmatization which in turn refer the words into original form to be compared directly using exact matching [17]. The exact matching of the lemmas across the two ontologies is performed. All matching pairs are considered as identical and involved in the alignment output.

The *non-words* strings are matched in two-ways comparison (see Figure 1.). The input strings are compared against the dictionary entities to extract all the possible corresponding dictionary entities with slight variations using an error-correction technique. The error correction searching procedure is taking place using a string similarity method such as edit-distance. Then, the candidates for each string are lemmatized in the same way as the words. Finally, the lemmatized candidates are compared against the words that are extracted directly in the first step and against each other.

3.1. Dictionary Look up

The dictionary look up is used for error detection in automatic error detection and correction applications for text which have been studied since 1960's [18]. The dictionary look up is a straight forward mechanism which clarifies whether specific words do exist in the dictionary. Given an input string, the dictionary looks up a word in the dictionary that has an exact match with the input string, it returns 'true' if such matching is found (the input string is a real word) and 'false' otherwise.

3.2. Error correction suggestions

There are several methods for error correction or spell suggestion for misspelled words. The developed methods and algorithm checks the misspelled words against the dictionary entities and retrieve the closer suggestions based on some similarity techniques. The very straight forward technique is edit-distance which is used in this paper with weight equal to 1, other techniques such as soundex, N-gram and probabilistic error corrections are used [18]. There are two forms of the retrieved suggestion(s): the best match and the multiple matches, we have used an order-multiple match's technique. Different preferable rate is given for the retrieved candidates based on their order. This technique limits the retrieved candidates to ten. We argue here that the proposed technique has the same ability as the exact matching, approximate matching and lexical based and outperform them in the error correction mechanism which can catch the typing errors, spelling errors and unlike the approximate string matching it finds firm correspondence. Moreover, the proposed approach does not require the use of threshold anymore which solves a difficult problem that is facing the approximate ontology alignment.

4. Implementation

We have implemented the proposed techniques using java, in Java JDK 1.6. Platform and using Integrated Development Environment: NetBeans IDE 5.5. The implemented solution has been developed to align OWL ontologies using Alignment API[1] [19]. The Alignment API providing tools for implementing, manipulation and evaluation of ontology alignment methods and approaches. 'Alignment API' is free java-based API that extends the OWL API[2]. The API has a variety of lexical and structural methods and is able to be extended with new methods. Beside, the Alignment API provides a variety of tools for manipulating alignment, comparing and generating variety of output forms. Figure 2 illustrates the implementation framework of the proposed technique.

[1] http://alignapi.gforge.inria.fr/
[2] http://owlapi.sourceforge.net/

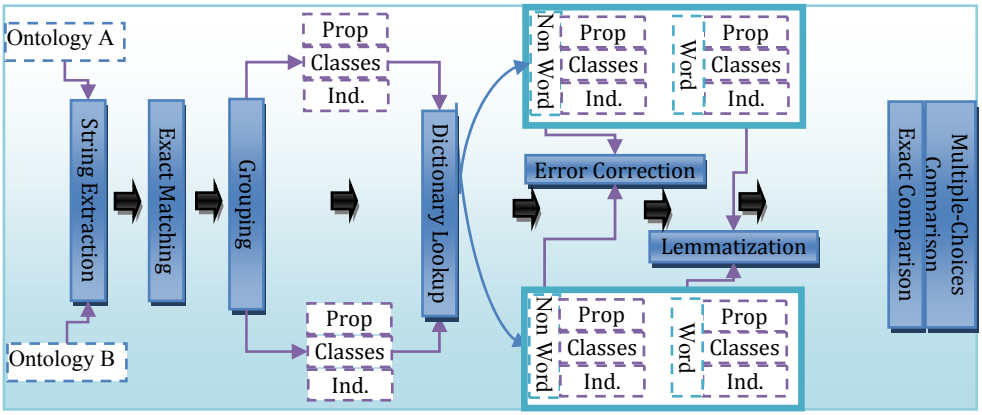

Fig. 2: Framework of the proposed technique

5. Result

The evaluation of the proposed technique is based on the experimental tests provided by first EON Ontology Alignment Contest[3] [17] which provide a set of variety tests in the domain of Bibliographic references. Single domain ontology, ontology 101, is used as an input to all the alignment tests, this ontology is aligned with a various other ontologies each has specific variation, these ontologies are: (102) irrelevant ontology from food domain. (103, 104): ontologies that differ from (101) in language generalization, no changes over the string properties. (201, 202): no names are used. (204): naming conventions using uppercase letter, underscores and dashes. (205): word synonyms are used. (206): foreign names in languages other than English. (221, 222,223,224,225,225,228): same string properties with different structure representation and eliminations. (230): expansion of classes' components and strings properties. (301, 302): Real ontologies for computer science in the bibliography domain. (303): real ontology with more items. (304): last real ontology by INRIA[4].

The true alignments for the mentioned tests are provided for the comparison purposes. The proposed alignment technique is evaluated using the well-known measures 'Precision' and 'Recall' which perform in the Alignment API. The 'precision' is the ratio between the true positive to the overall retrieved alignment. The 'Recall' is the ratio between the true positive to the total number of the true alignment exist and should be retrieved [20].

Table 1 represents the result of proposed technique and levenstein method. The results of levenstein method shown in the table are taken after applying 9 threshold values (0.1, 0.2, … , 0.9) and take the best results out of them.

As shown in the table the proposed technique has shown good results in some of the experimental tests, the points that can be highlighted about the proposed technique based on the results shown in Table 1, starting from test (102) there is no alignment between the two input ontologies, clearly because the inputs are two irrelevant ontologies. (103,104): Full recall, the input ontologies have same string properties for their elements. (201,202): no results because no names are

Table 1: Precision and Recall of Levenstein and the Developed Technique for the Alignment Tests

Test	Alignment Methods			
	Levenstein		Dictionary-based	
	Precision	Recall	Precision	Recall
101	1.0	1.0	1.0	1.0
102	Nan	Nan	Nan	Nan
103	0.97	0.97	0.97	1.0
104	0.99	0.99	0.99	1.0
201	Nan	Nan	Nan	Nan
202	Nan	Nan	Nan	Nan
204	0.93	0.93	0.96	0.78
205	0.6	0.32	0.79	0.3
206	Nan	Nan	Nan	Nan
221	1.0	1.0	1.0	1.0
222	0.93	0.93	0.97	0.94
223	0.93	0.93	0.93	0.96
224	0.99	0.99	0.99	1.0
225	0.99	0.99	0.99	1.0
228	1.0	1.0	1.0	1.0
230	0.87	0.97	0.97	0.97
301	0.9	0.7	0.95	0.57
302	0.95	0.64	0.95	0.78
303	0.87	0.81	0.95	0.94
304	0.97	0.94	0.97	.094

[3] http://oaei.ontologymatching.org/2004/Contest/
[4] http://www.inria.fr/index.en.html

used. (204): naming conventions, the proposed technique gives good precision, the low recall is due to using some short cuts which is not included in the dictionary such as 'MScthesis' with 'MasterThesis', 'TechReport' with 'TechnicalReport' which means that an enhancement over the proposed solution can be achieved using some enriched dictionary that includes such short cuts which are to be implemented in the future work. (205): very low recall because of using synonyms which is not handle in the proposed technique such as 'Proceeding' with 'inMinutes', the precision has gone down because the technique has matched strings that have similar single word and different stopped word such as 'periodicity' with 'inperiodical' the right alignment has to be 'frequency' with 'periodicity' which are synonyms. (206): other languages which is not handled in this technique. (221): full recall and precision. (222, 223): missing some matches that deals with definition of the words (synonyms) such as 'Book' with 'Reference', 'Article' with 'journalPart' and 'school' with 'higheducationinstitute'. (224,225,228): just an exact matching for exact strings. (230): the proposed technique resolve some matching difficulties such as matching 'Institution' with 'institutionName' , 'Organization' with 'organizationName' and 'Journal' with 'JournalName'. (301): missing synonyms such as (date) with 'hasYea' and resolve cases such as 'hasURL' with 'url' and 'hasChapter' with 'chapter'. (302): missing 'date' with 'PublishedOn' and 'LectureNotes' with 'Publication'. (303): resolved 'Event' with 'atEvent'. (304): accurate precision and resolve some matching such as 'Inproceeding' with 'proceeding' and 'inJournal' with 'Journal', missed those of synonyms such as 'Book' with 'inChapter'. We can conclude that the strength of the proposed technique is the ability to resolve the compound words and the compound words with word variations, and the major weakness is the ability to deal with word synonyms and short cuts which will be addressed using more lexical resources in the future work.

6. Future Work

As highlighted in the results. The precision of the proposed technique and the ability to resolve some difficult cases are good encouragements to proceed with the proposed technique. Using resources is a safe way to set matching and similarities between the Strings/Words which are of high value in the ontology. Dictionary and other lexical resources are extremely available for any purpose which allows developing more techniques that use such resources to ensure high precision, intelligence and safety matching. The lexical methods used normally as a first step in the hybrid systems for ontology alignment. The output of the lexical alignment used as initial reference pairs which used to set new matching based on some hierarchy similarity methods. Based on the obtained results the developed technique can be trusted to initiate such critical steps in hybrid systems.

7. Conclusion

An Ontology alignment based on a dictionary error correction has been developed. The proposed alignment is a combination of two most popular methods for lexical alignment which are the string-distance method and the semantic lexical method by using the error correction technique. We have argued that even non-word strings can be matched using this technique, because the error correction will give similar suggestions for similar strings. Thus, this solution provides a comprehensive mechanism for the lexical ontology alignment problem by taking into consideration that ontology is man made and their string properties forms using words non arbitrary strings. Moreover, the proposed approach does not require threshold anymore which solves a difficult problem that is facing the approximate ontology alignment.

8. Acknowledgment

The research presented is supported by a Research University grant titled 'Multimodal Meaning Normalization through Ontologies' (No: 1001/PKOMP/811021).

9. References

[1] Choi, N., I.-Y. Song, and H. Han, A Survey on ontology mapping. ACM SIGMOD Record, September 2006. 35: p. 34—41.

[2] Davide Fossati, G.G., Barbara Di Eugenio, Isabel Cruz, Huiyong Xiao, Rajen Subba, The problem of ontology alignment on the web: a first report, in 2nd Web as Corpus Workshop In conjunction with the 11th Conference of the European Chapter of the Association for Computational Linguistics Trento, Italy. 2006.

[3] Hage, W.R.v., S. Katrenko, and G.S. Schreiber, A Method to Combine Linguistic Ontology-Mapping Techniques, in Fourth International Semantic Web Conference 2005.

[4] Levenshtein, V., Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Soviet Physics-Doklady, 1966. 10: p. 707–710.

[5] MELICHAR, B. String matching with k differences by finite automata. in Proceedings of the International Congress on Pattern Recognition (ICPR '96). IEEE CS Press, Silver Spring, MD. 1996.

[6] Needleman, S.B., Wunsch, C. D, A general method applicable to the search for similarities in the amino acid sequences of two proteins. J. Mol. Biol, 1970. 48: p. 443-453.

[7] Myers, G., A fast bit-vector algorithm for approximate string matching based on dynamic programming. ACM, 1998. 46(3): p. 395–415.

[8] Winkler, W.E., The state of record linkage and current research problems. 1999.

[9] Stoilos, G., G. Stamou, and S. Kollias, A string metric for ontology alignment, in the 4th International Semantic Web Conference. 2005: Galway.

[10] Madhavan, J., P.A. Bernstein, and E. Rahm, Generic schema matching with Cupid., in 27th Intl. Conference on Very Large Databases (VLDB), . 2001: Rome, Italy. p. 49-58.

[11] Li, J. LOM: A Lexicon-based Ontology Mapping Tool. In Proceedings of the Performance Metrics for Intelligent Systems (PerMIS. '04). 2004.

[12] Miller, G.A., et al., Introduction to WordNet: An On-line Lexical Database. International Journal of Lexicography, 1990. 3(4): p. 235-244.

[13] Niles, I. and A. Pease, Toward a Standard Upper Ontology, in the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001). 2001.

[14] Pease, A., I. Niles, and J. Li, The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications, in In Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web. 2002.

[15] Niles, I. and A. Terry, The MILO: A general purpose, mid-level ontology, in International Conference on Information and Knowledge Engineering (IKE'04). 2004: Las Vegas, Nevada. p. 15-19.

[16] Keller, A.V.Z.a.U., Choosing an Ontology Language. Proceeding of World Academy of Science, engineering and technology. Vol (4)47-50, 2005.

[17] Sure, Y., et al. eds. in Proceedings of the 3rd International Workshop on Evaluation of Ontology-based Tools EON. 2004.

[18] Karen, K., Technique for automatically correcting words in text. ACM Comput. Surv., 1992. 24(4): p. 377-439.

[19] Euzenat, J. An API for Ontology Alignment. in 3ed conference on international semantic web conference (ISWC), . Hiroshima Japan. 2004

[20] Do, H.-H., S. Melnik, and E. Rahm. Comparison of Schema Matching Evaluations. in the 2nd Int. workshop on Web Databases. 2002.