Cross-Lingual Ontology Mapping and Its Use on the Multilingual Semantic Web

Bo Fu, Rob Brennan, Declan O'Sullivan
Knowledge and Data Engineering Group, School of Computer Science and Statistics,
Trinity College Dublin, College Green, Dublin 2, Ireland
{bofu, rob.brennan, declan.osullivan}@scss.tcd.ie

ABSTRACT
Ontology-based knowledge management systems enable the automatic discovery, sharing and reuse of structured data sources on the semantic web. With the emergence of multilingual ontologies, accessing knowledge across natural language barriers has become a pressing issue for the multilingual semantic web. In this paper, a semantic-oriented cross-lingual ontology mapping (SOCOM) framework is proposed to enhance interoperability of ontology-based systems that involve multilingual knowledge repositories. The contribution of cross-lingual ontology mapping is demonstrated in two use case scenarios. In addition, the notion of appropriate ontology label translation, as employed by the SOCOM framework, is examined in a cross-lingual ontology mapping experiment involving ontologies with a similar domain of interest but labelled in English and Chinese respectively. Preliminary evaluation results indicate the promise of the cross-lingual mapping approach used in the SOCOM framework, and suggest that the integrated appropriate ontology label translation mechanism is effective in the facilitation of monolingual matching techniques in cross-lingual ontology mapping scenarios.

Keywords
Cross-Lingual Ontology Mapping; Appropriate Ontology Label Translation; Matching Assessment Feedback; Querying of Multilingual Knowledge Repositories.

1. INTRODUCTION
The promise of the semantic web is that of a new way to organise, present and search information that is based on meaning and not just text. Ontologies are explicit and formal specifications of conceptualisations of domains of interests [11], thus are at the heart of semantic web technologies such as semantic search [8] and ontology-based information extraction [2]. As knowledge and knowledge representations are not restricted to the usage of a particular natural language, multilinguality is increasingly evident in ontologies as a result. Ontology-based applications therefore must be able to work with ontologies that are labelled in diverse natural languages. One way to realise this is by means of cross-lingual ontology mapping (CLOM).

In this paper, a summary of current CLOM approaches is presented in section 2. A semantic-oriented cross-lingual ontology mapping (SOCOM) framework that aims to facilitate mapping tasks carried out in multilingual environments is proposed and discussed in section 3. To illustrate possible applications of the SOCOM framework on the multilingual semantic web, two use case scenarios including cross-language document retrieval and personalised querying of multilingual knowledge repositories are presented in section 4. An overview of the initial implementation of the proposed framework is given in section 5. Section 6 presents an experiment that engages the integrated framework in a mapping scenario that involves ontologies labelled in English and Chinese, and discusses the evaluation results and findings from this experiment. Finally, work in progress is outlined in section 7.

WWW 2010, April 26-30, 2010, Raleigh, North Carolina, USA.

2. STATE OF THE ART
Current CLOM strategies can be grouped into five categories, namely manual processing, corpus-based approach, instance-based approach, linguistic enrichment of ontologies and the two-step generic approach. A costly *manual* CLOM process is documented in [13], where the English version of the AGROVOC [1] thesaurus is mapped to the Chinese Agriculture Thesaurus. Given large and complex ontologies, such an approach would be infeasible. Ngai et al. [16] propose a *corpus-based* approach to align the English thesaurus WordNet [2] and the Chinese thesaurus HowNet[3]. As bilingual corpora are not always available to domain-specific ontologies, it is difficult to apply their approach in practice. The *instance-based* approach proposed by Wang et al. [24] generates matching correspondences based on the analysis of instance similarities. It requires rich sets of instances embedded in ontologies, which is a condition that may not always be satisfied in the ontology development process. Pazienza & Stellato propose a *linguistically motivated* mapping method [17], advocating a linguistic-driven approach in the ontology development process that generates enriched ontologies with human-readable linguistic resources. To facilitate this linguistic enrichment process, a plug-in for the Protégé[4] editor – OntoLing [5] was also developed [18]. Linguistically enriched ontologies may offer strong evidence when generating matching correspondences. However, as such enrichment is not currently standardised, it is difficult to apply the proposed solution.

Trojahn et al. [23] present a multilingual ontology mapping framework, where ontology labels are first represented with collections of phrases in the target natural language. Matches are then generated using specialized monolingual matching agents that use various techniques (i.e. structured-based matching algorithms, lexicon-based matching algorithms and so on). However, as Shvaiko & Euzenat state in [20], "despite the many component matching solutions that have been developed so far, there is no integrated solution that is a clear success". Often various techniques are combined in order to generate high quality matching results [12], searching for globally accepted matches

[1] http://aims.fao.org/website/AGROVOC-Thesaurus/sub
[2] http://wordnet.princeton.edu
[3] http://www.keenage.com/html/e_index.html
[4] http://protege.stanford.edu
[5] http://art.uniroma2.it/software/OntoLing

can lead to a limited matching scope. In 2008, an OAEI[6] test case that involves the mapping of web directories written in English and Japanese was designed. Only one participant – the RiMOM tool – was able to submit results [26], by using a Japanese-English dictionary to translate labels from the Japanese web directory into English first, before applying monolingual matching procedures. This highlights the difficulty of exercising current monolingual matching techniques in CLOM scenarios.

Trojahn et al's framework and RiMOM's approach both employ a *generic two-step* method, where ontology labels are translated into the target natural language first and monolingual matching techniques are applied next. The translation process occurs in isolation of the mapping activity, and takes place independently of the semantics in the concerned ontologies. As a result, inadequate and/or synonymic translations can introduce "noise" into the subsequent matching step, where matches may be neglected by matching techniques that (solely) rely on the discovery of lexical similarities. This conception is further examined in [9], where strong evidence indicates that to enhance the performance of existing monolingual matching techniques in CLOM scenarios, appropriate ontology label translation is key to the generation of high quality matching results. This notion of selecting appropriate ontology label translations in the given mapping context is the focus of the SOCOM framework and the evaluation shown in this paper.

Notable work in the field of (semi-)automatic ontology label translation conducted by Espinoza et al. [7] introduces the LabelTranslator tool, which is designed to assist humans during the ontology localisation process. Upon selecting the labels of an ontology one at a time, ranked lists of suggested translations for each label are presented to the user. The user finally decides which suggested translation is the best one to localise the given ontology. In contrast to the LabelTranslator tool, the ontology rendition process of the SOCOM framework presented in this paper differs in its input, output and design purpose. Firstly, our rendering process takes formally defined ontologies (i.e. in RDF/OWL format) as input, but not the labels within an ontology. Secondly, it outputs formally defined ontologies labelled in the target natural language, but not lists of ranked translation suggestions. Lastly, our rendering process is designed to facilitate further machine processing (more precisely, existing monolingual ontology matching techniques), whereas the LabelTranslator tool aims to assist humans.

3. THE SOCOM FRAMEWORK

Given ontologies O_1 and O_2 (see Figure 1) that are labelled in different natural languages, O_1 is first transformed by the SOCOM framework into an equivalent of itself through the *ontology rendering* process as O_1'. O_1' contains all the original semantics of O_1 but is labelled in the natural language that is used by O_2. O_1' is then matched to O_2 using *monolingual matchers* to generate candidate matches, which are then reviewed by the *matching assessment* mechanism in order to establish the final mappings.

Ontology renditions are achieved by structuring the translated ontology labels in the same way as the original ontology O_1, and assigning these translation labels to new namespaces to create well-formed resource URIs in O_1' (for more details, please see [9]). Note that the structure of O_1 is not changed during this process, as Giunchiglia et al. [10] point out, the conceptualisation of a particular ontology node is captured by its label and its position in the ontology structure. Thus, the ontology rendering process should not modify the position of a node, because doing so would effectively alter the semantics of the original ontology.

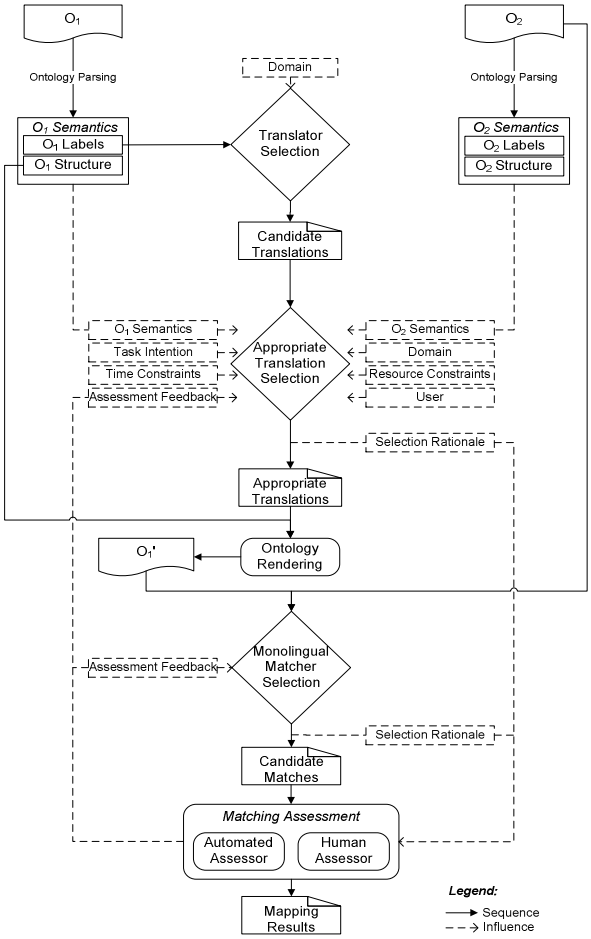

Figure 1. SOCOM Framework Workflow Overview

In contrast to the generic approach, where the translation of ontology labels takes place in isolation from the ontologies concerned, the SOCOM framework is semantic-oriented and aims to identify the most appropriate translation for a given label. To achieve this, firstly, suitable translation tools are selected at the *translator selection* point to generate candidate translations. This selection process is influenced by the knowledge domain of the concerned ontologies. For general knowledge representations, off-the-shelf machine translation (MT) tools or thesauri can be applied. For specific domains such as the medical field, specialised translation media are more appropriate. Secondly, to identify the most appropriate translation for a label among its candidate translations, the *appropriate translation selection* process is performed. This selection process is under the influence of several information sources including the source ontology semantics, the target ontology semantics, the mapping intent, the operating domain, the time constraints, the resource constraints, the user and finally the matching assessment result feedback. These influences are explained next.

The *semantics defined in O_1* can indicate the context that a to-be-translated label is used in. Given a certain position of the node with this label, the labels of its surrounding nodes (referred to as surrounding semantics in this paper) can be retrieved and studied. For example, for a class node, its surrounding semantics can be

[6] http://oaei.ontologymatching.org

represented by the labels of its super/sub/sibling-classes. For a property node, its surrounding semantics can be represented by the labels of the resources which this property restricts. For an individual, the surrounding semantics can be characterised by the label of the class it belongs to. Depending on the granularity of the given ontologies in a mapping scenario, an ontological resource's surrounding semantics should be modelled with flexibility. For example, if the ontologies are rich in structure, immediate surrounding resource labels (e.g. direct super/sub relations) alone can form the content of the surrounding semantics. If the ontologies are rich in instance, where the immediate surrounding label (e.g. the class an instance belongs to) alone is weak to provide the instance's context of use, indirect (e.g. all super/sub classes declared in the ontology) resource labels should be included in the surrounding semantics. The goal of obtaining surrounding semantics of a given resource is to provide the translation selection process with additional indications of the context a resource is used in[7].

As O_1 is transformed so that it can be best mapped to O_2, the *semantics defined in* O_2 therefore can act as broad translation selection rules. When several translation candidates are all linguistically correct for a label in O_1, the most appropriate translation is the one that is most semantically similar to what is used in O_2. An example of appropriate ontology label translation is shown in Figure 2, where the source ontology is labelled in Chinese and is mapped to an English target ontology. The class *摘要* from the source ontology has translation candidates *abstract* and *summary*. To determine the most appropriate translation, the defined semantics of the target ontology can influence the translation selection process. To understand how this is possible, consider three scenarios. Figure 2a demonstrates a situation where a class named *Summary* exists in the target ontology. In this case, *Summary* would be considered as more appropriate than *abstract* since it is the exact label used by the target ontology. Figure 2b illustrates another scenario where the target ontology contains a class named *Sum*. From a thesaurus or a dictionary, one can learn that *Sum* is a synonym of *summary*, therefore, instead of using either *abstract* or *summary*, *Sum* will be chosen as the appropriate translation in this case. Figure 2c shows a third scenario where both *Abstract* and *Summary* exist in the target ontology, the appropriate translation is then concluded by studying the surrounding semantics. The source class *摘要* has a super-class *出版物* (with translation candidates *publication* and *printing*), two sibling-classes *章节* (with translation candidates *chapter* and *section*) and *书籍* (with translation candidates *book* and *literature*). Its surrounding semantics therefore include: {*publication*, *printing*, *chapter*, *section*, *book*, *literature*}. Similarly, in the target ontology, the surrounding semantics of the class *Summary* contains: {*BookChapter*, *Reference*}, and the surrounding semantics of the class *Abstract* would include: {*Mathematics*, *Applied*}. Using string comparison techniques, one can determine that the strings in the surroundings of the target class *Summary* are more similar to those of the source class. *Summary* therefore would be the appropriate translation in such a case. Note that the SOCOM framework is concerned with searching for appropriate translations (from a mapping point of view) but not necessarily the most linguistically correct translations (from a natural language processing point of view), because our motivation for translating ontology labels is so that the ontologies can be best mapped[8]. This should not be confused with translating labels for the purpose of ontology localisation, where labels of an ontology are translated so that it is "adapted to a particular language and culture" [21].

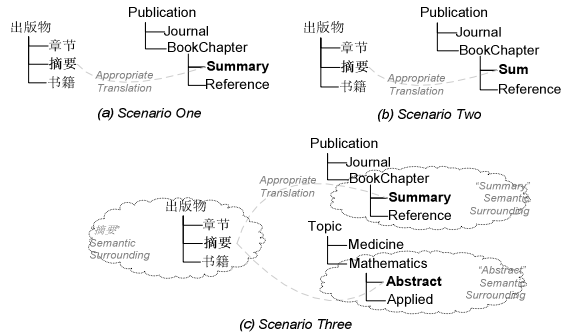

Figure 2. Examples of Appropriate Label Translation

In addition to using the embedded semantics of the given ontologies, *task intention* can also influence the outcome of the translation selection process as it captures some of the mapping motives. Consider a CLOM scenario where the user is not comfortable with all the natural languages involved, and would like to test just how meaningful/useful it is to map the given ontologies. In such a case, the selection of translation candidates need not be very sophisticated, thus results returned from off-the-shelf MT tools can be acceptable. The *domain* of the ontologies is another influence on the translation selection process. For example, if O_1 and O_2 are domain representations where each one is associated with collections of documents in different natural languages, lists of frequently used words in these documents can be collected. The translation candidate that is ranked highest on these lists would be deemed as the most appropriate translation. Moreover, *time constraints* can influence the translation selection process. If the mappings must be conducted dynamically such as the work presented in [5], the translation selection consequently must be fast, where it might not make use of all the resources that are available to it. On the other hand, not all of the aforementioned resources will be available in every CLOM scenario. *Resource constraints* therefore can have an impact on the outcome of the translation selection process. Furthermore, *users*, at times, can have the expertise that is not obtained by the system, and should influence the translation selection process when necessary. Lastly, *matching result feedback* can influence the future selection of appropriate translations (discussed next).

[7] The generation of surrounding semantics presented in this paper does not attempt to estimate the semantic relatedness between concepts, it is a procedure performed within readily defined ontologies in a cross-lingual ontology mapping scenario that aims to gather the context of use for a particular resource in the given ontologies. Though one might assume that the SOCOM framework would work best when ontologies with similar granularity are presented, this however, is not a requirement of the framework. As already mentioned, the surrounding semantics are modelled with flexibility, where indirectly related concepts in the ontology would be collected as long as the surrounding well illustrates the context of use for a particular ontological resource.

[8] Note that the appropriate ontology label translation mechanism presented in this paper does not attempt to disambiguate word senses, as the appropriateness of a translation is highly restricted to the specific mapping scenarios, thus it is not a form of natural language processing technique.

Once O_1' is generated, various monolingual matching techniques can be applied to create matches between O_1' and O_2. The selection of these monolingual matchers depends on the feedback generated from the mapping result assessment. *Assessment feedback* can be implicit (i.e. pseudo feedback) or explicit. Pseudo feedback is obtained automatically, where the system assumes matches that meet certain criteria are correct. For example, "correct" results may be assumed to be the ones that have confidence levels of at least 0.5. The precision of the matches generated can then be calculated for each matching algorithm used, which will allow the ranking of these algorithms. The ranking of the MT sources can also be determined upon establishment of the usage of each MT source (i.e. as percentages) among the "correct" matches. Based on these rankings, the top performing MT tools and matching algorithms can then be selected for the future executions of the SOCOM framework. Explicit feedback is generated from users and is more reliable than pseudo feedback, which can aid the mapping process in the same way as discussed above.

Matching assessment feedback allows insights into how the correct mappings are generated, in particular, which translation tool(s) and matching algorithm(s) are most suitable in the specified CLOM scenario. Such feedback in turn could influence the future selection of appropriate label translations and the monolingual matching techniques to use. Finally, the feedback should be influenced by the *selection rationale* employed during the translation selection process and the monolingual matching process. Such rationale can be captured as metadata as part of the mapping process and include information such as the influence sources used, translation tools used, monolingual matching techniques used, similarity measures of semantic surroundings and so on. The use of matching assessment feedback addresses one of the scalability issues that arise. Consider a mapping scenario where the concerned ontologies contain thousands of entities, one way to rapidly generate mapping results and improve mapping quality dynamically is to use the pseudo feedback. For the first, e.g. 100 mapping tasks, assume the ones that satisfy certain criteria are correct, detect how they are generated, and keep using the same techniques for the remaining mapping tasks. This assessment process can also be recursive where the system is adjusted for every few mapping tasks. Finally, explicit feedback involves users in the mapping process, which contributes towards addressing one of the challenges, namely user involvement in ontology matching as identified by Shvaiko & Euzenat in [20].

4. USE CASES

The notion of using conceptual frameworks such as thesauri and ontologies in search systems [6] [4] for improved information access [19] and enhanced user experiences [22] is well researched in the information retrieval (IR) and the cross-lingual IR (CLIR) community. However, the use of ontology mapping as a technique to aid the search functions in IR has been relatively limited. The most advanced work of using ontology alignment in CLIR, to the best of our knowledge, is Zhang et al.'s statistical approach presented in [25], which does not involve translations of ontology labels. To avail statistical analysis such as latent semantic indexing, singular value decomposition, directed acyclic graphs and maximal common subgraph on document collections, parallel corpora must be generated beforehand. However, this often is an expensive requirement and may not always be satisfied. Also, by applying statistical techniques only, such an approach ignores the existing semantic knowledge within the given ontologies in a mapping scenario. Hence alternative solutions are in need. The SOCOM framework presented in this paper can contribute towards this need. Its contribution can be demonstrated through two use cases as shown in Figures 3 & 4.

User generated content such as forums often contain discussions on how to solve particular technical problems, and a large amount of content of this type is written in English. Consider a scenario illustrated in Figure 3, where the user whose preferred natural language is Portuguese is searching for help on a forum site, but the query in Portuguese is returning no satisfactory results. Let us assume that the user also speaks English as a second language and would like to receive relevant documents that are written in English instead. To achieve this, domain ontologies in Portuguese and English can be extracted based on text presented in the documents using such as Alani et al.'s approach [1]. Mappings can then be generated pre-runtime using the SOCOM framework between the Portuguese ontology and the English ontology, and stored as RDF triples. At run time, once a query is issued in Portuguese, it is first transformed using such as Lopez et al.'s method [14] to associate itself with a concept in the Portuguese domain ontology. This Portuguese concept's corresponding English concept(s) can then be obtained by looking it up in the mapping triplestore. Once the system establishes which English concepts to explore further, their associated documents in English can be retrieved.

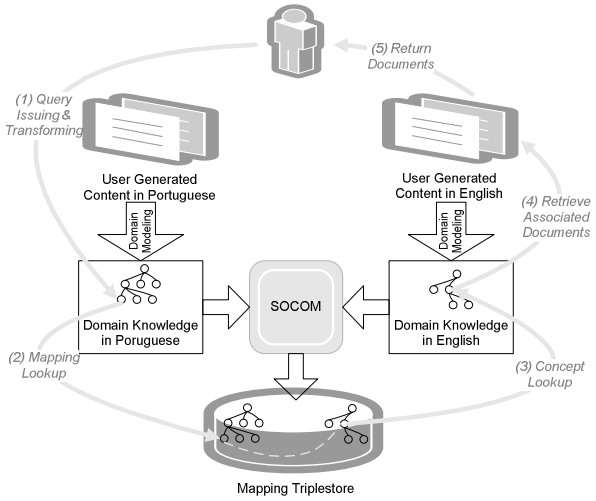

Figure 3. SOCOM Enabled Cross-Language Document Retrieval

Personalisation can also be enhanced with the integration of the SOCOM framework in scenarios such as the one shown in Figure 4, where a user is bi/multi-lingual and would like to receive documents in a restricted knowledge domain in various natural languages as long as they are relevant. To achieve this, ontology-based user models[9] containing knowledge such as user interests and language preferences can be generated pre-runtime using approaches such as [3]. Similar to the previous scenario, domain ontologies labelled in different natural languages can be obtained from sets of documents. In Figure 4, knowledge representations in English, French, German and Spanish are obtained in ontological form. Mappings of the user model and the various domain ontologies can then be generated using the

[9] User modelling is a well researched area particularly in adaptive hypermedia and personalised search systems, however, this is outside the scope of this paper.

SOCOM framework. At run time, a user query is transformed to be associated with a concept or concepts in the user model. By looking up in the mapping triplestore, the matched concepts in various knowledge repositories (the German and the Spanish knowledge repositories in the case of Figure 4) can be obtained, which will then lead to the retrieval of relevant documents in different natural languages.

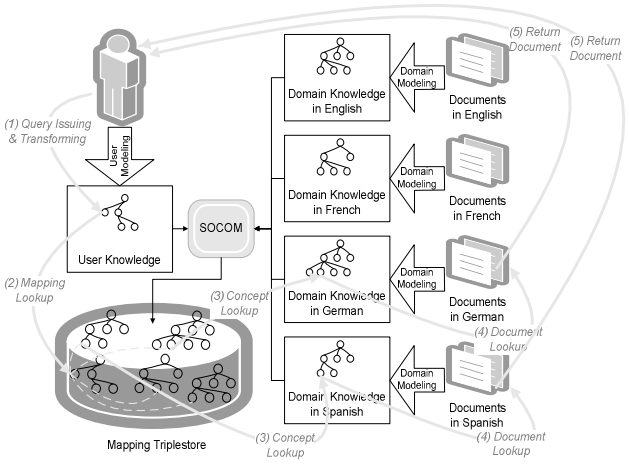

Figure 4. Personalised Querying of Multilingual Knowledge Repositories with SOCOM

5. IMPLEMENTATION

To examine the soundness of the appropriate ontology label translation selection process proposed in the SOCOM framework, an initial implementation of the proposal has been completed that uses just the semantics within the given ontologies in a CLOM scenario. This light-weight translation selection process (i.e. one that includes semantics in O_1 and semantics in O_2, but excludes the six other influence sources as shown in Figure 1) is the focus of the implementation and the evaluation presented in this paper.

This initial SOCOM implementation integrates the Jena 2.5.5 Framework[10] to parse the formally defined input ontologies. To collect candidate translations for ontology labels in O_1, the GoogleTranslate[11] 0.5 API and the WindowsLive[12] translator are used[13]. Synonyms of ontology labels in O_2 are generated by querying WordNet[14] 2.0 via the RiTa[15] API. Ontology labels are often concatenated to create well-formed URIs (as white spaces are not allowed), e.g. a concept *associate professor* can be labelled as *AssociateProfessor* in the ontology. As the integrated MT tools cannot process such concatenated labels, they are split into sequences of their constituent words before being passed to the MT tools. This is achieved by recognising concatenation patterns. In the previous example, white spaces are inserted before each capital letter found other than the first one. The candidate translations are stored in a translation repository, whereas the synonyms are stored in a lexicon repository. Both repositories are stored in the eXist[16] 1.0rc database.

The appropriate translation selection process invokes the repositories in the database via the XML:DB[17] 1.0 API, to compare each candidate translation of a given source label to what is stored in the lexicon repository. An overview of this appropriate translation selection process can be seen in Figure 5. If a one-to-one match (note that the match found in the lexicon repository can be either a target label used in O_2, or a synonym of a target label that is used in O_2) is found, the (matched target label or the matched synonym's corresponding) target label is selected as the appropriate translation. If one-to-many matches (i.e. when several target labels and/or synonyms in the lexicon repository are matched) are found, the surrounding semantics (see section 3) of the matched target labels are collected and compared to the surrounding semantics of the source label in question. Using a space/case-insensitive edit distance string comparison algorithm based on Nerbonne et al.'s method [15], the target label with surrounding semantics that are most similar to those of the source resource is chosen as the most appropriate translation. If no match is found in the lexicon repository, for each candidate translation, a set of interpretative keywords are generated to illustrate the meaning of this candidate. This is achieved by querying Wikipedia[18] via the Yahoo Term Extraction Tool[19]. Using the same customised string comparison algorithm, the candidate with keywords that are most similar to the source label's surrounding semantics is deemed as the most appropriate translation.

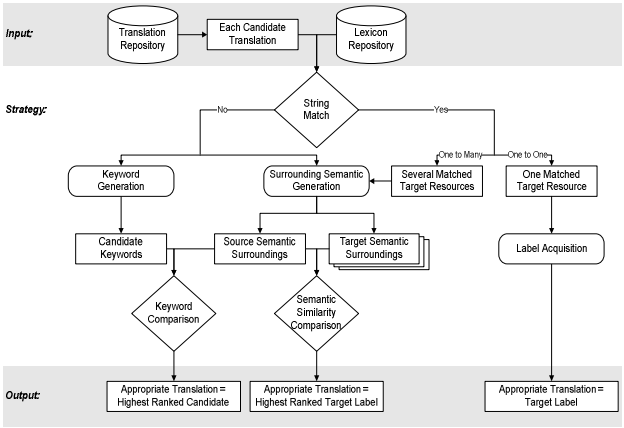

Figure 5. Overview of the Appropriate Ontology Label Translation Selection Process

Once appropriate translations are identified for each label in O_1, given the original source ontology structure, O_1' is generated using the Jena Framework. Finally, O_1' is matched to O_2 to generate candidate matches via the Alignment API[20] version 3.6.

6. EVALUATION

To evaluate the effectiveness of the integrated appropriate translation selection process, this initial implementation of the SOCOM framework is engaged in a CLOM experiment that

[10] http://jena.sourceforge.net
[11] http://code.google.com/p/google-api-translate-java
[12] http://www.windowslivetranslator.com/Default.aspx
[13] One could use a dictionary/thesaurus here, however, as the appropriate ontology label translation selection process in the SOCOM framework is not a word sense disambiguation mechanism (see section 3), off-the-self MT tools are efficient to collect candidate translations.
[14] http://wordnet.princeton.edu
[15] http://www.rednoise.org/rita
[16] http://exist.sourceforge.net
[17] http://xmldb-org.sourceforge.net/index.html
[18] http://www.wikipedia.org
[19] http://developer.yahoo.com/search/content/V1/termExtraction.html
[20] http://alignapi.gforge.inria.fr

involves ontologies labelled in Chinese and English describing the research community domain, against a baseline system – the generic approach, where labels are translated in isolation using just the GoogleTranslate 0.5 API and matches are generated using the Alignment API[21] version 3.6 (see [9] for more technical details of the implementation of the generic approach).

6.1 Experimental Setup

Figure 6 gives an overview of the experiment. A Chinese ontology CSWRC[22] is created manually by a group of domain experts (excluding the authors of this paper) based on the English SWRC[23] ontology. It contains 54 classes, 44 object properties and 30 data type properties. This Chinese ontology is matched to the English ISWC[24] ontology (containing 33 classes, 18 object properties, 17 data type properties and 50 instances) using the generic approach and the SOCOM approach, generating results M-G and M-S respectively.

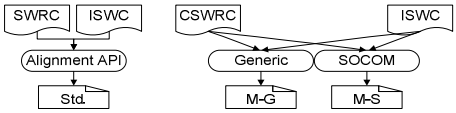

Figure 6. Cross-Lingual Ontology Mapping Experiments

As the CSWRC ontology is formally and semantically equivalent (with the same structured concepts but labelled in Chinese) to the SWRC ontology, a reliable set of gold standard (referred to as Std. in Figure 6) can be generated as matches found between the SWRC ontology and the ISWC ontology using the Alignment API[25]. By comparing results M-G and M-S to Std., this experimental design aims to find out which approach can generate higher quality matching results, when the concerned ontologies hold distinct natural languages and varied structures.

6.2 Experimental Results

Precision and recall[26] scores of M-G and M-S are calculated, see Figure 7, where a match is considered correct as long as the identified pair of corresponding resources is included in the gold standard Std., regardless of its confidence level.

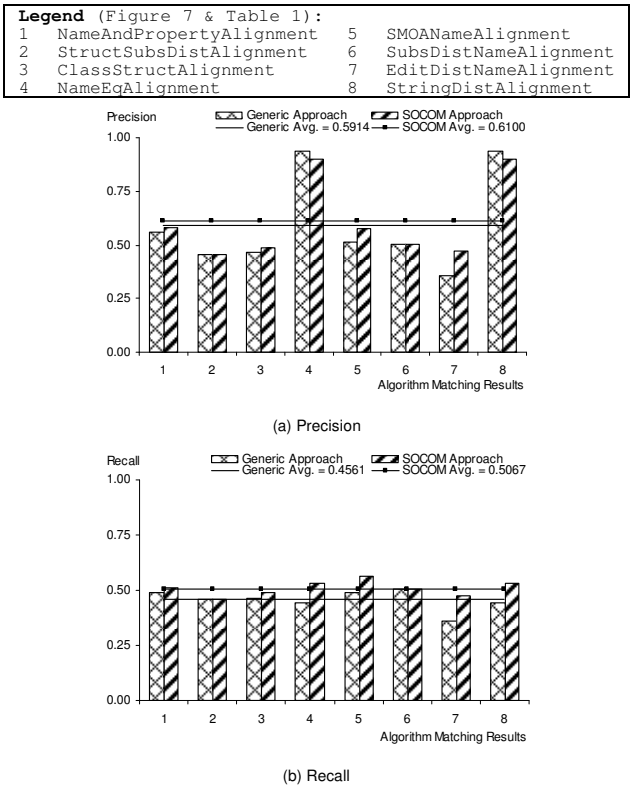

Figure 7. Overview of Precision and Recall when Disregarding Confidence Levels

Figure 7a shows that except the NameEqAlignment and the StringDistAlignment algorithm, all other matching methods indicate equal or higher precision when using the SOCOM approach. The aforementioned two algorithms employ strict string comparison techniques, where no dissimilarity between two labels is overlooked. Though this is a desirable characteristic at times, in this particular experiment setting, some matches are neglected in Std.. E.g. when using the StringDistAlignment algorithm, the gold standard was unable to establish a match between the class *AssociateProfessor* (in SWRC) and the class *Associate_ Professor* (in ISWC) because these labels are not identical, although this would have been a sound match if a human was involved or if preprocessing was undertaken. When the SOCOM approach is used to match CSWRC to ISWC, the most appropriate translation for the class *副教授* (associate professor) in the source ontology was determined as *Associate_Professor* since this exact English label was used in the target ontology. Consequently, a match with 1.00 confidence level between the two was generated in M-S. However, as this correspondence was not included in Std., such a result is deemed as incorrect. Similar circumstances led to the lower precision scores of the SOCOM approaches in cases that involve the NameEqAlignment and the StringDistAlignment algorithms. Nevertheless, on average, with a precision score at 0.61, the SOCOM approach generated more correct matching results than the generic approach overall. Furthermore, at an average recall score of 0.5067 (see Figure 7b), the SOCOM approach demonstrates that its correct results are always more complete than those generated by the generic approach.

As precision and recall each measures one aspect of the match quality, f-measure scores are calculated to indicate the overall

---

[21] The Alignment API 3.6 contains eight matching algorithms, namely NameAndPropertyAlignment, StructSubsDistAlignment, ClassStructAlignment, NameEqAlignment, SMOANameAlignment, SubsDistNameAlignment, EditDistNameAlignment and StringDistAlignment. For each correspondence found, a matching relationship is given and is accompanied by a confidence measure that range between 0 (not confident) and 1 (confident).

[22] http://www.scss.tcd.ie/~bofu/SOCOMExperimentJuly2009/Ontologies/CSWRC.owl

[23] http://ontoware.org/frs/download.php/298/swrc_v0.3.owl

[24] http://annotation.semanticweb.org/iswc/iswc.owl

[25] Based on the assumption that the CSWRC ontology is equivalent to the SWRC ontology, this experimental design aims to validate whether matches generated using the exact same matching algorithms would result the same or highly similar corresponding concepts.

[26] Given a gold standard with R number of matching results, and an evaluation set containing X number of results, if N number of them are correct based on the gold standard, then for this evaluation set precision = N/X, recall = N/R and f-meaure = 2/(1/precision + 1/recall).

quality[27]. Table 1 shows that the SOCOM approach generated results with at least equal quality compared to the generic approach. In fact, the majority of algorithms were able to generate higher quality matches when using the SOCOM approach, leading to an average of 0.5460 in its f-measure score. The differences in the two approaches' f-measure scores (when they exist) range from a smallest 1.9% (when using the NameAndPropertyAlignment algorithm) to a highest of 11.4% (when using the EditDistNameAlignment algorithm). Additionally, when using the SOCOM approach, bigger differences in f-measure can be seen in lexicon-based algorithms. Such a finding indicates that appropriate ontology label translation in the SOCOM framework contributes positively to the enhanced performances of matching algorithms, particularly those that are lexicon-based.

Table 1. **F-measure Scores when Disregarding Confidence Levels**

	Generic	SOCOM
1	.5233	.5421
2	.4574	.4574
3	.4651	.4884
4	.6000	.6667
5	.5020	.5714
6	.5039	.5039
7	.3571	.4714
8	.6000	.6667
Avg.	.5011	.5460

So far, the confidence levels of matching results have not been taken into account. To include this aspect in the evaluation, confidence means of the correct matches and their standard deviations are calculated. The mean is the average confidence of the correct matches found in a set of matching results, where the higher it is, the better the results. The standard deviation is a measure of dispersion, where the greater it is, the greater the spread in the confidence levels. Higher quality matching results therefore are those with higher means and lower standard deviations. On average, when using the SOCOM framework, the confidence mean is 0.7105. Whereas, a lower mean of 0.6970 is found in the generic approach. The standard deviation when using the SOCOM framework is 0.2134, which is lower than 0.2161 as found in the generic approach. These findings denote that matches generated using the SOCOM approach are of higher quality, because they are not only more confident but also less dispersed.

Moreover, average precision, recall and f-measure scores are collected at various thresholds. These scores are calculated when the conditions a correct result must satisfy adjust, i.e. a matching result is only considered correct when it is included in the gold standard, and it has confidence level of at least 0.25, 0.50, 0.75 or 1.00. An overview of the trends is shown in Figure 8. As the requirement for a correct matching result become stricter, the precision (Figure 8a) and recall (Figure 8b) scores both decline as a result, leading to a similar decreasing trend in the f-measure (Figure 8c) scores. The differences in the recall scores of the two approaches are greater than the differences of their precision scores. This finding suggests that the matches generated using the two approaches may appear similar in their correctness, but the ones generated by the SOCOM approach are more complete. Overall, the SOCOM approach always has higher precision, recall and f-measure scores than the generic approach no matter what the threshold is[28]. This finding further confirms that the matches generated using the SOCOM approach are of higher quality.

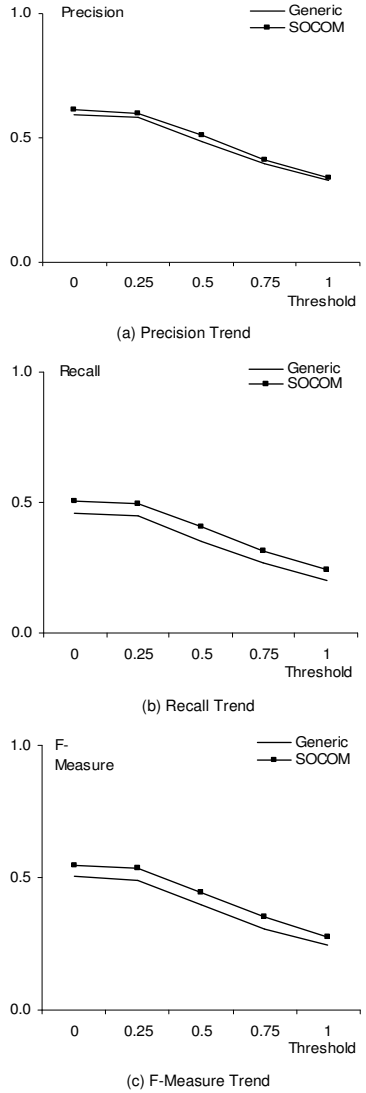

Figure 8. Trend Overview in Average Precision, Recall and F-Measure

Lastly, one can argue that the differences in the f-measure scores found between the generic and the SOCOM approach are rather small and therefore can be ignored. To validate the difference (if it exists) of the two approaches, paired t-tests are carried out on the f-measure scores collected across various thresholds, and a p-value of 0.001 is found. At a significance level of α=0.05, it can be concluded that the f-measure scores are statistically significant, meaning that the SOCOM approach generated higher quality matches than the generic approach.

[27] Note that neither precision nor recall alone is a measurement of the overall quality of a set of matching results, as the former is a measure for correctness and the latter is a measure for completeness. One can be sacrificed for the optimisation of the other, for example, when operating in the medical domain, recall may be sacrificed in order to achieve high precision; when merging ontologies, the opposite may be desired.

[28] Dotted lines of the generic and the SOCOM approach shown in Figure 8 are almost parallel to one another, this may be in part a result of the engineering approach deployed in the experiment (i.e. using the same tools in the implementation for both approaches). Further research, however, is needed to confirm the validity of this speculation.

7. CONCLUSIONS & FUTURE WORK

A semantic-oriented framework to cross-lingual ontology mapping is presented and evaluated in this paper. Preliminary evaluation results of an early prototype implementation illustrate the effectiveness of the integrated appropriate ontology label translation mechanism, and denote a promising outlook for applying CLOM techniques in multilingual ontology-based applications. The findings also suggest that a fully implemented SOCOM framework – i.e. one that integrates all the influence factors (discussed in section 2) – would be even more effective in the generation of high quality matches in CLOM scenarios.

The implementation of such a comprehensive SOCOM framework is currently on-going. It is planned to be evaluated using the benchmark datasets from the OAEI 2009 campaign, engaging the proposed framework in the mapping of ontologies that are written in very similar natural languages, namely English and French. In addition, the SOCOM framework is to be embedded in a demonstrator cross-language document retrieval system as part of the Centre for Next Generation Localisation, which involves several Irish academic institutions and a consortium of multi-national industrial partners aiming to develop novel localisation techniques for commercial applications.

8. ACKNOWLEDGMENT

This research is partially supported by Science Foundation Ireland (Grant 07/CE/11142) as part of the Centre for Next Generation Localisation (http://www.cngl.ie) at Trinity College Dublin.

9. REFERENCES

[1] Alani H., Kim S., Millard D. E., Weal M. J., Hall W., Lewis P. H., Shadbolt N. R.. Automatic ontology-based knowledge extraction from Web documents. IEEE Intelligent Systems 18, 1, 14-21, Jan. 2003

[2] Buitelaar P., Cimiano P., Frank A., Hartung M., Racioppa S.. Ontology-based information extraction and integration from heterogeneous data sources. International Journal of Human Computer Studies, 66, 11, 759-788, Nov. 2008

[3] Cantador I., Fernández M., Vallet D., Castells P., Picault J., Ribière M.. A multi-purpose ontology-based approach for personalised content filtering and retrieval. Advances in Semantic Media Adaptation and Personalization. Studies in Computational Intelligence, vol. 93, 25-51, 2008

[4] Castells P., Fernández M., Vallet D.. An adaptation of the vector-Space model for ontology-based information retrieval. IEEE Transactions on Knowledge and Data Engineering 19(2), Special Issue on Knowledge and Data Engineering in the Semantic Web Era, 261-272, Feb. 2007

[5] Conroy C., Brennan R., O'Sullivan D., Lewis D.. User evaluation study of a tagging approach to semantic mapping. In Proceedings of ESWC, 623-637, 2009

[6] De Luca E. W., Eul M., Nürnberger A.. Multilingual query-reformulation using an RDF-OWL EuroWordNet representation. In Proceedings of the Workshop on Improving Web Retrieval for Non-English Queries (iNEWS07), at SIGIR 2007, ISBN 978-84-690-6978-3, 55-61, 2007

[7] Espinoza M., Gómez-Pérez A., Mena E.. LabelTranslator – a tool to automatically localize an ontology. In Proceedings of ESWC, 792-796, 2008

[8] Fernandez M., Lopez V., Sabou M., Uren V., Vallet D., Motta E., Castells P.. Semantic search meets the Web. In Proceedings of IEEE ICSC, 253-260, 2008

[9] Fu B., Brennan R., O'Sullivan D.. Cross-lingual ontology mapping – an investigation of the impact of machine translation. In Proceedings of ASWC, LNCS 5926, 1-15, 2009

[10] Giunchiglia F., Yatskevich M., Shvaiko P.. Semantic matching: algorithms and implementation. Journal on Data Semantics, vol. IX, 1-38, 2007

[11] Gruber T.. A translation approach to portable ontologies. Knowledge Acquisition 5(2):199-220, 1993

[12] Li J., Tang J., Li Y., Luo Q.. RiMOM: A dynamic multistrategy ontology alignment framework. IEEE Transactions on Knowledge and Data Engineering, Vol. 21, No. 8, 1218-1232, 2009

[13] Liang A. C., Sini M.. Mapping AGROVOC and the Chinese agricultural thesaurus: definitions, tools, procedures. New Review of Hypermedia and Multimedia, 12:1, 51-62, 2006

[14] Lopez V., Uren V., Motta E., Pasin M.. AquaLog: an ontology-driven question answering system for organizational semantic intranets. Web Semantics. 5, 2, 72-105, Jun. 2007

[15] Nerbonne J., Heeringa W., Kleiweg P.. Edit distance and dialect proximity. Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison, 2nd ed. CSLI, Stanford, v-xv, 1999

[16] Ngai G., Carpuat M., Fung P.. Identifying concepts across languages: a first step towards a corpus-based approach to automatic ontology alignment. In Proceedings of the 19th International Conference on Computational Linguistics, vol.1, 1-7, 2002

[17] Pazienta M., Stellato A.. Linguistically motivated ontology mapping for the Semantic Web. In Proceedings of the 2nd Italian Semantic Web Workshop, 14-16, 2005

[18] Pazienza M. T., Stellato A.. Exploiting linguistic resources for building linguistically motivated ontologies in the Semantic Web. In Proceedings of OntoLex Workshop, 2006

[19] Shuang L., Fang L., Clement Y., Weiyi M.. An effective approach to document retrieval via utilizing WordNet and recognizing phrases. 27th Annual international ACM SIGIR Conference on Research and Development in information Retrieval, 266-272, ACM Press, 2004

[20] Shvaiko P., Euzenat J.. Ten challenges for ontology matching. In Proceedings of ODBASE, 1164-1182, 2008

[21] Suárez-Figueroa M. C., Gómez-Pérez A.. First attempt towards a standard glossary of ontology engineering terminology. In Proceedings of the 8th International Conference on Terminology and Knowledge Engineering (TKE'08), 2008

[22] Stamou, S., Ntoulas, A.. Search personalization through query and page topical analysis. User Modeling and User-Adapted Interaction 19, 1-2, 5-33., Feb. 2009

[23] Trojahn C., Quaresma P., Vieira R.. A framework for multi-lingual ontology mapping. In Proceedings of LREC, 1034-1037, 2008

[24] Wang S., Englebienne G., Schlobach S.. Learning concept mappings from instance similarity. In Proceedings of ISWC, 339-355, 2008

[25] Zhang L., Wu G., Xu Y., Li W., Zhong Y.. Multilingual collection retrieving via ontology alignment. In Proceeding of ICADL 2004, LNCS 3334, 510-514, Springer-Verlag, 2004

[26] Zhang X., Zhong Q., Li J., Tang J., Xie G., Li H.. RiMOM results for OAEI 2008. In Proceedings of the OM Workshop, 182-189, 2008