Cross-lingual Ontology Mapping – An Investigation of the Impact of Machine Translation

Bo Fu, Rob Brennan, and Declan O'Sullivan

Centre for Next Generation Localisation & Knowledge and Data Engineering Group,
School of Computer Science and Statistics, Trinity College Dublin, Ireland
{bofu, rob.brennan, declan.osullivan}@cs.tcd.ie

Abstract. Ontologies are at the heart of knowledge management and make use of information that is not only written in English but also in many other natural languages. In order to enable knowledge discovery, sharing and reuse of these multilingual ontologies, it is necessary to support ontology mapping despite natural language barriers. This paper examines the soundness of a generic approach that involves machine translation tools and monolingual ontology matching techniques in cross-lingual ontology mapping scenarios. In particular, experimental results collected from case studies which engage mappings of independent ontologies that are labeled in English and Chinese are presented. Based on findings derived from these studies, limitations of this generic approach are discussed. It is shown with evidence that appropriate translations of conceptual labels in ontologies are of crucial importance when applying monolingual matching techniques in cross-lingual ontology mapping. Finally, to address the identified challenges, a semantic-oriented cross-lingual ontology mapping (SOCOM) framework is proposed and discussed.

Keywords: Cross-lingual Ontology Mapping; Multilingual Ontologies; Ontology Rendering.

1 Introduction

The evolution of the World Wide Web in recent years has brought innovation in technology that encourages information sharing and user collaboration as seen in popular applications during the Web 2.0 era. The future of the Web – the Semantic Web will "allow for integration of data-oriented applications as well as document-oriented applications" [1]. In the process of achieving this goal, ontologies have become a core technology for representing structured knowledge as well as an instrument to enhance the quality of information retrieval [2] [3] and machine translation [4]. Benjamins et al [5] identify multilinguality as one of the six challenges for the Semantic Web, and propose solutions at the ontology level, annotation level and the interface level. At the ontology level, support should be provided for ontology engineers to create knowledge representations in diverse native natural languages. At the annotation level, tools should be developed to aid the users in the annotation of ontologies regardless of the natural languages used in the given ontologies. Finally, at

the interface level, users should be able to access information in natural languages of their choice. This paper aims to tackle challenges at the annotation level, in particular, it investigates issues involved in cross-lingual ontology mapping and aims to provide the necessary support for ontology mapping in cross-lingual environments. *Cross-lingual ontology mapping (CLOM) refers to the process of establishing relationships among ontological resources from two or more independent ontologies where each ontology is labeled in a different natural language.* The term multilingual ontologies in this paper refers to independent ontologies *o* and *o´* where the labels in *o* are written in a natural language which is different from that of the labels in *o´*. It must not be confused with representing concepts in one ontology using multilingual labels. In addition, this paper focuses on multilingual ontologies that have not been linguistically enriched, and are specified according to the Resource Description Framework (RDF) schema[1]. Furthermore, this paper presents a first step towards achieving CLOM in generic knowledge domains, which can be improved upon to accommodate more sophisticated CLOM mapping strategies among ontologies in more refined, particular knowledge domains.

A generic approach is investigated in this paper, CLOM is achieved by first translating the labels of a source ontology into the target natural language using freely available machine translation (MT) tools, then applying monolingual ontology matching techniques to the translated source ontology and the target ontology in order to establish matching relationships. In particular, the impact of MT tools is investigated and it is shown with evidence that when using the generic approach in CLOM, the quality of matching results is dependent upon the quality of ontology label translations. Based on this conclusion, a semantic-oriented cross-lingual ontology mapping (SOCOM) framework is proposed which is specifically designed to map multilingual ontologies and to reduce noise introduced by MT tools. The remainder of this paper is organised as follows, section 2 discusses related work. Section 3 details the application of the aforementioned generic approach in CLOM experiments which involve mappings of ontologies labeled in Chinese and English. Findings and conclusions from these experiments are presented and discussed in section 4. The proposed SOCOM framework and its current development are discussed in section 5.

2 Related Work

Considered as light weight ontologies, thesauri often contain large collections of associated words. According to the Global WordNet Association[2], (at the time of this publication) there are over forty WordNet[3]-like thesauri in the world covering nearly 50 different natural languages, and counting. Natural languages used include Arabic (used in ArabicWordNet[4]); Bulgarian (used in BulNet[5]); Chinese (used in HowNet[6]);

[1] http://www.w3.org/TR/rdf-schema
[2] http://www.globalwordnet.org
[3] http://wordnet.princeton.edu
[4] http://www.globalwordnet.org/AWN
[5] http://dcl.bas.bg/BulNet/general_en.html

Dutch, French, German, Italian, Spanish (used in EuroWordNet[7]); Irish (used in LSG[8]) and many others. To make use of such enormous knowledge bases, research has been conducted in the field of thesaurus merging. This is explored when Carpuat et al [6] merged thesauri that were written in English and Chinese into one bilingual thesaurus in order to minimize repetitive work while building ontologies containing multilingual resources. A language-independent, corpus based approach was employed to merge WordNet and HowNet by aligning synsets from the former and definitions of the latter. Similar research was conducted in [7] to match Dutch thesauri to WordNet by using a bilingual dictionary, and concluded a methodology for vocabulary alignment of thesauri written in different natural languages. Automatic bilingual thesaurus construction with an English-Japanese dictionary is presented in [8], where hierarchies of words can be generated based on related words' co-occurrence frequencies. Multilinguality is not only found in thesauri but also evident in RDF/OWL ontologies. For instance, the OntoSelect Ontology Library[9] reports that more than 25% (at the time of this publication) of its indexed 1530 ontologies are labeled in natural languages other than English[10]. To enable knowledge discovery, sharing and reuse, ontology matching must be able to operate across natural language barriers. Although there is already a well-established field of research in monolingual ontology matching tools and techniques [9], as ontology mapping can no longer be limited to monolingual environments, tools and techniques must be developed to assist mappings in cross-lingual scenarios.

One approach of facilitating knowledge sharing among diverse natural languages builds on the notion of enriching ontologies with linguistic resources. A framework is proposed in [10] which aims to support the linguistic enrichment process of ontological concepts during ontology development. A tool – OntoLing[11] is developed as a plug-in for the ontology editor Protégé[12] to realise such a process as discussed in [11]. Similar research aiming to provide multilingual information to ontologies is discussed in [12], where a linguistic information repository is proposed to link ontological concepts with lexical resources. Such enrichment of ontologies provide knowledge engineers with rich linguistic data and can be used in CLOM, however, in order for computer-based applications to make use of these data, standardisation of the enrichment is required. As such requirement is currently not included in the OWL 2 specification[13], it would be difficult to make use of the vast number of monolingual ontology matching techniques that already exist.

Similar to linguistically enriching ontologies, translating the natural language content in ontologies is another approach to enable knowledge sharing and reuse. The translation of the multilingual AGROVOC thesaurus[14] is discussed in [13], which

---

6 http://www.keenage.com
7 http://www.illc.uva.nl/EuroWordNet
8 http://borel.slu.edu/lsg
9 http://olp.dfki.de/ontoselect
10 http://olp.dfki.de/ontoselect;jsessionid=3B72F3160F4D7592EE3A5CCF702AAE00?wicket:bookmarkablePage=:de.dfki.ontoselect.Statistics
11 http://art.uniroma2.it/software/OntoLing
12 http://protege.stanford.edu
13 http://www.w3.org/TR/owl2-profiles
14 http://aims.fao.org/en/website/AGROVOC-Thesaurus/sub

involves a large amount of manual work and proves to be time and human resource consuming. An ontology label translation tool, LabelTranslator is demonstrated in [14]. It is designed to provide end-users with ranked translation suggestions for ontology labels. The motivation of its design is to ensure that information represented in an ontology using one particular natural language could still achieve the same level of knowledge expressivity if translated into another natural language. Users must select labels to be translated one at a time, LabelTranslator then returns the selected label's suggested translations in one of the three target natural languages, English, Spanish and German. It can be used to provide assistance in the linguistic enrichment process of ontologies as discussed in [15]. LabelTranslator is designed to assist the human to perform semi-automatic ontology label translations and linguistic enrichments, it is not concerned with generations of machine-readable ontologies in the target natural language so that matching tools can manipulate. In contrast to LabelTranslator, the ontology rendering process presented in this paper differs in its input, output and aim. Firstly, the input of our ontology rendering process is ontologies and not ontology labels. Secondly, the output of this rendering process is machine-readable formally defined ontologies that can be manipulated by computer-based systems such as monolingual matching tools. Lastly, such an ontology rendering design aims to facilitate CLOM, it is designed to assist further machine processing whereas the LabelTranslator tool aims to assist humans.

An example of CLOM scenario is illustrated by the Ontology Alignment Evaluation Initiative[15] (OAEI) contest in 2008, where a test case requiring the mapping of web directories written in English and Japanese was defined[16]. Among thirteen participants, only four took part in this test scenario with results submitted from just one contestant. Zhang et al. [16] used a dictionary to translate Japanese words into English (it is unclear whether this translation process is manual or automated) before carrying out the matching process using RiMOM. The generic approach presented in this paper is based on Zhang et al.'s method, instead of using a dictionary, freely available MT tools are used. Montiel-Ponsoda & Peters [17] classify three levels to localizing multilingual ontologies, at the terminological layer, at the conceptual layer and at the pragmatic layer. The translation process presented in the generic CLOM approach concerns translations at the terminological layer, i.e., the terms used to define classes and properties are translated into the target natural language. Pazienza & Stellato propose a linguistically motivated approach to ontology mapping in [18]. The approach urges the usage of linguistically enriched expressions when building ontologies, and envisions systems that can automatically discover the embedded linguistic evidence and establish alignments that support users to generate sound ontology mapping relationships. However, as mentioned previously, the multilingual linguistically enriched ontologies demanded by this approach are hard to come by when such specifications are not currently included in the OWL 2 standardization effort. Trojahn et al. propose a multilingual ontology mapping framework in [19], which consists of smart agents that are responsible for ontology translation and capable of negotiating mapping results. For each ontology label, the translation agent looks up a dictionary and returns a collection of results in the target

[15] http://oaei.ontologymatching.org
[16] http://ri-www.nii.ac.jp/OAEI/2008

natural language. The ontology labels are then represented with a group of the returned translation results. Once source and target ontologies are in the same natural language, they are passed to the mapping process which consists of three types of mapping agents, lexical, semantic and structural. These agents each conclude a set of mapping results with an extended value-based argumentation algorithm. Finally, globally accepted results are generated as the final set of mappings [20]. Such an approach is based on the assumption that correct mapping results are and always will be generated by various matching techniques regardless of the algorithms used. However, as stated by Shvaiko & Euzenat [21], "despite the many component matching solutions that have been developed so far, there is no integrated solution that is a clear success", therefore, looking for globally accepted results may limit the scope of correct mapping relationship discovery. In contrast, the proposed SOCOM framework in this paper aims to maximize the performance of individual monolingual ontology matching algorithms in CLOM by providing them with ontology renditions that contain appropriate label translations.

3 A Generic Approach to Cross-lingual Ontology Mapping

A generic approach to achieve CLOM is presented in this section, as shown in figure 1. Given two ontologies representing knowledge in different natural languages, the ontology rendering process first creates a translated source ontology which is an equivalent of the original source ontology, only labeled in the target natural language. Then monolingual matching tools are applied to generate matching results between the translated source ontology and the target ontology. An integration of the generic approach is discussed in section 3.1. To evaluate the soundness of this approach, two experiments involving the Semantic Web for Research Communities (SWRC) ontology[17] and the ISWC ontology[18] were designed to examine the impact of MT tools in the process of ontology rendering (discussed in section 3.2), also the quality of matching results generated using such an approach (discussed in section 3.3).

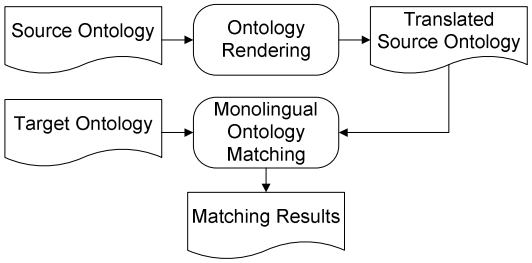

Fig 1. A Generic Cross-lingual Ontology Mapping Approach

[17] http://ontoware.org/frs/download.php/298/swrc_v0.3.owl
[18] http://annotation.semanticweb.org/ontologies/iswc.owl

3.1 Integration of the Generic Approach

The ontology rendering process shown in figure 1 is achieved with a Java application – OntLocalizer, which generates machine-readable, formally defined ontologies in the target natural language by translating labels of the given ontology's concepts using MT tools, assigning them with new namespaces and structuring these resources – now labeled in the target natural language – using the Jena Framework[19] in the exact same way as the original ontology. Figure 2 shows the components of the OntLocalizer tool. Labels of ontology resources are extracted first by the Jena Framework, which are then passed onto the MT tools to generate translations in the target natural language. Given the original ontology's structure, these translated labels can be structured accordingly to create the translated source ontology. The integrated MT tools include the GoogleTranslate API[20] and the SDL FreeTranslation[21] online translator.

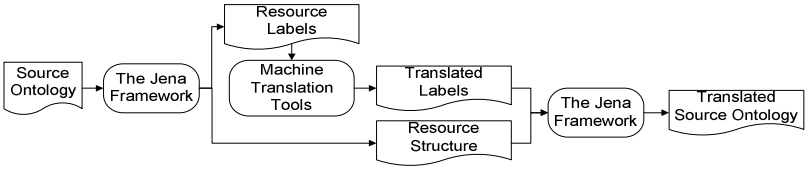

Fig 2. OntLocalizer Component Overview

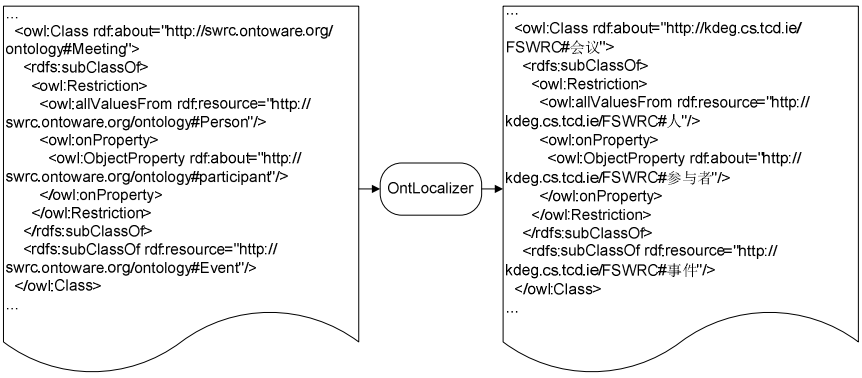

Fig 3. An Example of Ontology Translation

As white spaces are not allowed in the naming of the ontological resources, ontology labels often contain phrases that are made up by two or more words. An example of such labels can be a class named "AssistantProfessor", where the white space between two words has been removed and capital letters are used to indicate the beginning of another word. Another example can be an object property labeled as "is_about", where the white space between two words has been replaced by an underscore. As these labels cannot be translated by the integrated MT tools, the

[19] http://jena.sourceforge.net
[20] http://code.google.com/p/google-api-translate-java
[21] http://www.freetranslation.com

OntLocalizer tool thus breaks up such labels to sequences of constituent words based on the composing pattern, before sending them to the MT tools. In the aforementioned examples, "AssistantProfessor" is transformed to "Assistant Professor", and "is_about" is transformed to "is about". Now both in their natural language forms, phrases "Assistant Professor" and "is about" are passed to the MT tools to generate results in the target natural language. Such a procedure is not required when translating labels written in languages such as Chinese, Japanese etc., as phrases written in these languages naturally do not contain white spaces between words and can be processed by the integrated MT tools. Finally, when structuring the translated labels, white spaces are removed to create well-formed resource URIs. *Translation collisions* can happen when a translator returns the same result for several resources in an ontology. For instance, in the SWRC ontology, using the GoogleTranslate API (version 0.4), the class "Conference" and the class "Meeting" are both translated into "会议" (meaning "meeting" in Chinese). To differentiate the two, the OntLocalizer tool checks whether such a resource already exists in the translated source ontology. If so, a number is assigned to the resource label which is under consideration. In the aforementioned example, "Conference" becomes "会议" and "Meeting" becomes "会议 0" in the translated ontology. As the integrated MT tools only return one translation result for each intake phrase, it is therefore unnecessary to disambiguate the returned translations in the experiment. A part of the SWRC ontology and its translation in Chinese using the OntLocalizer tool is shown in figure 3.

Once the source ontology is labeled in the target natural language, monolingual ontology matching techniques can be used to generate matching results. Currently, this is achieved by the Alignment API[22] (version 2.5).

3.2 Experiment One Design and Integration

Experiment one is designed to examine the impact of MT tools in the process of ontology rendition, specifically, the quality of machine translated resource labels. In this experiment, labels in the SWRC ontology are translated from English to Chinese through two media, the OntLocalizer tool and a human domain expert – being the lead author. Three translated versions of the SWRC ontology are then created, the GSWRC ontology when using the GoogleTranslate API, the FSWRC ontology when using the FreeTranslation online translator, and the HSWRC ontology which is created manually using the Protégé ontology editor. Each translated version has the original structure of the SWRC ontology with new namespaces assigned to labels in the target natural language. The SWRC ontology is mapped to itself to generate a gold standard of the matching results as M(1), which consists of pairs of matched ontology resources in English. M(A) which contains results of matched resources in Chinese, is then created when the HSWRC ontology is mapped to itself. If exactly the same pairs of resources are matched in M(A) as those found in M(1), then M(A) can be considered as the gold standard in Chinese. The GSWRC ontology and the FSWRC ontology are then each mapped to the HSWRC ontology to create the mappings M(B)

[22] http://alignapi.gforge.inria.fr

and M(C), both containing matched resources in Chinese. Finally, M(B) and M(C) are compared against M(A). This process is shown in figure 4. Eight matching algorithms supported by the Alignment API are used in this experiment.

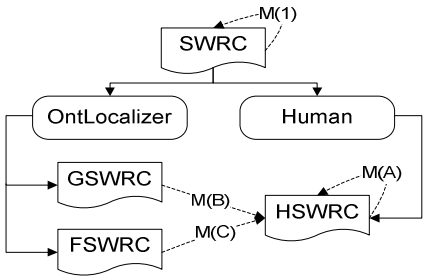

Fig 4. Experiment One Overview

The hypothesis of this experimental setup is to verify whether the label translation procedure using MT tools would impact on the quality of translated ontologies. If M(B) and M(C) show the same set of results as suggested by M(A), it would mean that MT tools are able to perform like humans and a generic approach using them in CLOM is ideal. If M(B) and M(C) proves to be poorly generated, it would mean that the ontology rendition process is flawed.

3.2 Experiment Two Design and Integration

The second experiment is designed to further investigate the impact of MT tools in CLOM by evaluating the quality of matching results generated using the generic approach. An overview of the experimental steps is shown in figure 5. The English SWRC ontology and the English ISWC ontology are both translated by OntLocalizer to create ontologies labeled in Chinese. The GSWRC ontology and the GISWC ontology are created when using the GoogleTranslate API, and the FSWRC ontology and the FISWC ontology are generated when using the SDL FreeTranslation online translator integrated in OntLocalizer.

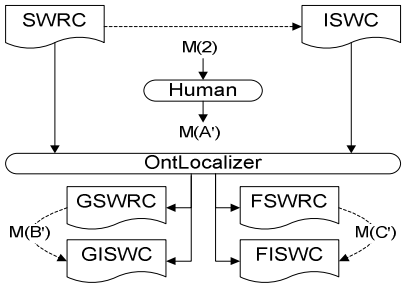

Fig 5. Experiment Two Overview

The original SWRC ontology is mapped to the original ISWC ontology to generate M(2) as the gold standard which contains matched resources in English.

M(B') is generated when the GSWRC ontology is mapped to the GISWC ontology, similarly M(C') is generated when the FSWRC ontology is mapped to the FISWC ontology. Both M(B') and M(C') contain matched resources in Chinese. Again eight matching algorithms provided by the Alignment API were used in every mapping. To evaluate the quality of M(B') and M(C'), they are compared against the gold standard. Since M(2) contains matched resources written in English, the labels of these resources are translated manually to Chinese by the lead author as M(A'). M(A') is then regarded as the gold standard. Evaluations of M(B') and M(C') are finally conducted based on comparisons to M(A'). The hypothesis of this experiment is, if M(B') and M(C') generated the same sets of matching results as M(A'), it would mean that the generic approach is satisfactory to achieve CLOM. If M(B'), M(C') fail to conclude the same results as found in the gold standard, it would mean that the generic approach would be error-prone when applied to CLOM scenarios.

Precision, recall, fallout and f-measure scores were calculated in both experiments for all the matching algorithms used. Precision measures the correctness of a set of results. Recall measures the completeness of the number of correct results. Fallout measures the number of incorrect matching results based on the gold standard. Finally, f-measure can be considered as a determination for the overall quality of a set of results. If the established gold standard has R number of results and a matching algorithm finds X number of results, among which N number of them are correct according to the gold standard, then precision = N/X; recall = N/R; fallout = (X-N)/X; and f-measure = 2/(1/precision + 1/recall). All scores range between 0 and 1, with 1 being very good and 0 being very poor. An example can be that low fallout score accompanied by high precision and recall scores denote superior matching results.

4 Findings and Conclusions

Findings and conclusions from the two experiments are presented in this section. The results of experiment one is presented and discussed in section 4.1. Section 4.2 shows the results from the second experiment. Finally, data analysis is given in section 4.3.

4.1 Experiment One Findings

Regardless of the matching algorithms used from the Alignment API, the exact same sets of matching results generated in M(1) were found in M(A). Thus, it is with confidence that M(A) can be considered as the gold standard in Chinese. Figure 5 shows an overview of the evaluation results of experiment one. As M(A) equals M(1), its precision, recall and f-measure scores are 1.00 and with 0.00 fallout. The results generated by the eight matching algorithms from the Alignment API are evaluated based on comparisons made to M(A). In M(B) and M(C), a pair of matched resources is considered correct when it is found in the gold standard regardless of its confidence level. Such an evaluation approach aims to measure the maximum precision, recall and f-measure scores that can be achieved in the generated results.

As figure 5 shows, in experiment one, *NameEqAligment* and *StringDistAlignment* algorithm had the highest precision score, however, their low recall scores resulted just above the average f-measure scores. Structure-based matching algorithms had lower recall scores and higher fallout scores comparing to lexicon-based matching algorithms. For each set of results evaluated, the precision score is always higher than its other scores, which suggests that a considerable number of correct matching results is found, however, they are always incomplete. On average, regardless of the matching algorithms used, f-measure scores are almost always less than 0.50, showing that none of the matching algorithms could meet the standard which is set by the gold standard. Moreover, M(B)'s average f-measure is 0.4272, whereas M(C)'s average f-measure is 0.3992, which suggests that GoogleTranslate API performed slightly better than SDL FreeTranslation online translator in this experiment. Nevertheless, it must be noted that neither of the MT tools was able to generate a translated ontology which, when mapped to itself, could produce a same set of results that are determined by the gold standard. This finding suggests that MT tools had a negative impact on the quality of ontology rendition output.

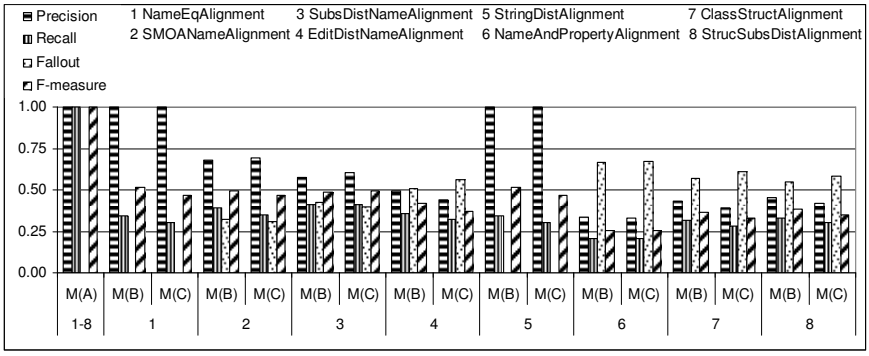

Fig 5. Experiment One Results

4.2 Experiment Two Findings

To further validate this finding, the same evaluation approach is used in the second experiment, where a pair of matched result is considered correct as long as it is found in the gold standard, regardless of its confidence level. A series of gold standards were generated for each of the eight matching algorithms in M(2) – written in English, and later manually translated as M(A') – written in Chinese. The evaluation of the results found in M(B') and M(C') is shown in Figure 6.

The *StringDistAlignment* matching algorithm had the highest precision and recall scores in this experiment, thus yielding the highest f-measure score in M(B') and M(C'). Similar to the results found in experiment one, structure-based matching algorithms had lower recall scores comparing to lexicon-based matching algorithms. In experiment two, fallout scores for all the matching algorithms are higher than that of experiment one's, which suggests that the matching procedure was further complicated by the translated ontologies. Also, f-measure scores indicate that structure-based matching algorithms were unable to perform as well as lexicon-based

matching algorithms. The average f-measure in M(B') was 0.2927 and 0.3054 in M(C'), which suggests that the FreeTranslation online translator had a slightly better performance than the Google Translate API in this experiment. Nevertheless, from an ontology matching point of view, such low f-measure scores would mean that when used in CLOM, the generic approach would only yield less than fifty percent of the correct matching results. The findings from experiment two show that it is difficult for matching algorithms to maintain high-quality performance when labels have been translated in isolation using MT tools, and the generic approach in CLOM can only yield poor matching results.

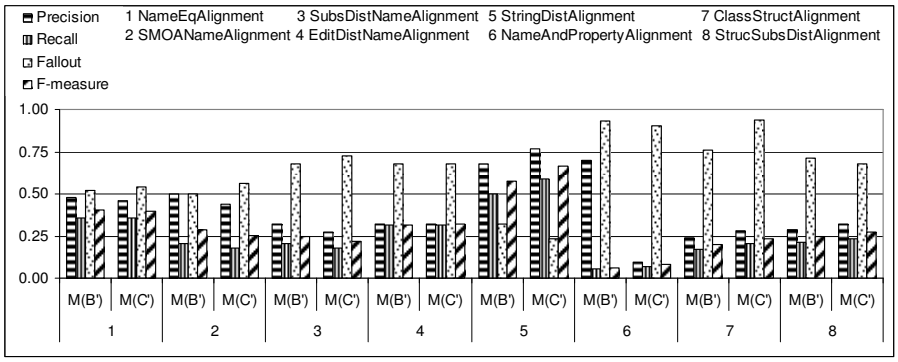

Figure 6. Experiment Two Results

4.3 Result Analysis

So far, the evaluation results that are shown in the previous sections disregard confidence levels. When these confidence levels are taken into account, it is shown that there is a drop in the number of matching results generated with absolute confidence. Table 1 gives an overview of the percentages of matching results with 1.00 confidence levels. In both experiments, all pairs of matched resources generated by the *NameEqAlignment* algorithm and the *NameAndPropertyAlignment* algorithm have 1.00 confidence levels. This is not the case for other algorithms however, where more than half of the results with absolute confidence was not found. For example, every matched pairs of resources by the *EditDistNameAlignment* algorithm from the gold standard in experiment one had 1.00 confidence levels. This was not achieved in M(B) or M(C), where the former contained 47.31% of confident results and only 41.94% for the latter. Averagely, the gold standard in experiment one established a 96.25% of confident results, whereas only 49.53% were found in M(B) and 49.37% in M(C). A similar finding can be concluded for experiment two based on the statistics shown in table 1.

Findings from the experiments suggest that if automated MT tools are to be used in CLOM, more specifically, in the ontology rendering process, the quality of translated ontologies needs to be improved in order for monolingual matching tools to generate high quality matching results. Translation errors introduced by the MT tools in the experiments can be categorized into three main categories. *Inadequate*

translation – as mentioned earlier in section 3.1, "Conference" and "Meeting" were both translated into the same words in Chinese. However, since conference is a specified type of meeting, the translated term was not precise enough to capture the intended concept presented in the original ontology. This can be improved if given the context of a resource label to be translated, i.e. the context of a resource can be indicated by a collection of associated property labels, super/sub-class labels. *Synonymic translation* – where the translation result of a label is correct, however it is different with the one that was used by the target ontology. This can be accounted by algorithms that take structural approaches when establishing matching results, however, it can be very difficult for lexicon-based algorithms to associate them. This can be improved if several candidates are provided in the translation process, and the selection of these candidates gives priority to labels which are used by the target ontology. *Incorrect translation* – where the translation of a term is simply wrong, yielding poor matching results. Similar to inadequate translations, this can be improved if the context of an ontology resource is known to the translation process.

Table 1. Matched Pairs of Results with 1.00 Confidence Levels (%)

	1	2	3	4	5	6	7	8	Avg.
M(A)	*100.00*	*77.34*	*100.00*	*100.00*	*100.00*	*92.68*	*100.00*	*100.00*	*96.25*
M(B)	100.00	33.78	47.83	47.31	100.00	37.25	15.05	15.05	49.53
M(C)	100.00	35.38	44.32	41.94	100.00	34.62	19.35	19.35	49.37
M(A')	*100.00*	*30.89*	*26.56*	*48.57*	*100.00*	*30.36*	*0.00*	*10.94*	*43.42*
M(B')	100.00	16.00	30.86	36.23	100.00	11.63	3.23	3.23	37.65
M(C')	100.00	18.00	30.59	38.24	100.00	13.95	1.30	4.30	38.30

1 = NameEqAlignment
2 = SMOANameAlignment
3 = SubsDistNameAlignment
4 = EditDistNameAlignment
5 = StringDistAlignment
6 = NameAndPropertyAlignment
7 = ClassStructAlignment
8 = StrucSubsDistAlignment

To overcome these challenges and maximise the performance of monolingual matching tools in CLOM, *appropriate translations* of ontology labels must be achieved. A Semantic-Oriented Cross-lingual Ontology Mapping (SOCOM) framework designed to achieve this is proposed and discussed in the next section.

5 The SOCOM Framework and On-going Research

The semantic-oriented cross-lingual ontology mapping (SOCOM) framework is presented and discussed in this section. The SOCOM framework illustrates a process that is designed specifically to achieve CLOM, it has an extensible architecture that aims to accommodate easy integrations of off-the-shelf software components. To address challenges identified in the experiments and reduce noise introduced by the MT tools, the selection of appropriate translated labels is under the influence of labels used in the target ontology. The SOCOM framework divides the mapping task into three phases – an ontology rendering phase, an ontology matching phase and a matching audit phase. The first phase of the SOCOM framework is concerned with the rendition of an ontology labeled in the target natural language, particularly,

appropriate translations of its labels. The second phase concerns the generation of matching results in a monolingual environment. Finally, the third phase of the framework aids ontology engineers in the process of establishing accurate and confident mapping results. Ontology matching is the identification of candidate matches between ontologies, whereas ontology mapping is the establishment of the actual correspondence between ontology resources based on candidate matches [22], this distinction is reflected in the SOCOM framework. Figure 7 shows a process diagram of the proposed framework.

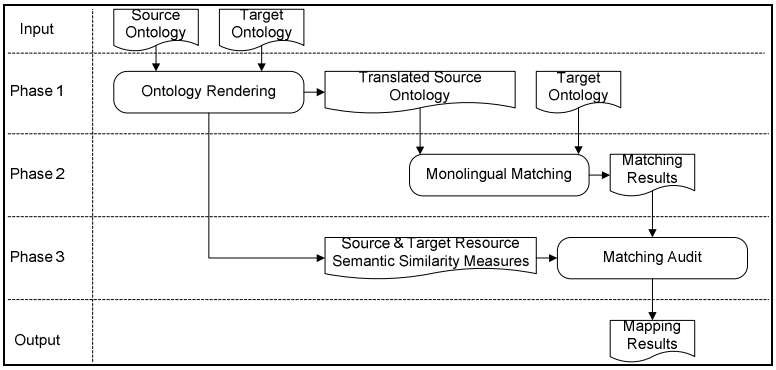

Fig 7. The SOCOM Framework Process Diagram

In phase one, the SOCOM framework searches for the most appropriate translation results for ontology labels in the target natural language. To achieve this, the selection of translation candidates is defined by the context a resource is used in, and influenced by the labels that appear in the target ontology. As experimental results show that translating ontology labels in isolation leads to poorly translated ontologies which then yields low-quality matching results, thus, label translations should be conducted within context. As the meaning of a word vary depending on the context it is used in, it is therefore important to capture what a word/phrase signifies as accurately as possible in the target natural language. For instance, the sentence *there is a shift in the tone of today's news broadcasts* and the sentence *research shows that an inevitable side effect of night shifts is weight gain* both use the word *shift*. However, in the first sentence, it is used to express *a change*, whereas in the second sentence, it refers to *a period of work*. In the SOCOM framework, to capture the meaning of a word/phrase in the ontology rendering phase, the context is characterised by the surrounding ontology concepts. As the purpose of translating the source ontology is so that it can be mapped to the target ontology for generations of high quality mapping results (i.e. the translation of the source ontology concepts is within a specific context), the identification of the most appropriate translation results is aided by the labels that appear in the target ontology. Instead of blindly accepting translation results that are returned from a MT tool, for each resource label, a group of translation results are collected and treated as translation candidates. A *translation repository* containing source labels and their translation candidates can be created given a source ontology. On the other hand, a *lexicon repository* can be constructed based on the labels presented in a given target ontology. For each target label, a

collection of synonyms can be assembled to maximize knowledge representation with various words and phrases other than those that originally appeared in the target ontology. This can be achieved by querying dictionaries, WordNet, etc., or accessing refined lexicon bases for precise knowledge domains with strict vocabularies such as medicine. Each of the candidates can then be compared to the phrases in the lexicon repository. When matches are found with a target label or a target label's synonym, the target label is chosen as the most appropriate translation result. In addition, when translations are compared to terms in the lexicon repository, similarity measures can be calculated using string comparison techniques, which can then assist the ontology engineers in the final mapping process.

In the second phase, as the source ontology is now labeled in the target natural language, the SOCOM framework can apply existing monolingual ontology matching techniques. It is assumed that prior to CLOM using the SOCOM framework, human experts are involved to establish that it is meaningful to map the concerned ontologies, i.e. they cover the same/similar domain of interest, they are reliable, complete and similar in granularity.

Lastly, in phase three, the matching audit procedure aids ontology engineers in the process of generating the final mapping results. This procedure makes use of the semantic similarity measures that have been concluded in phase one, and displays these findings to the mapping expert providing background information to assist the final mapping. Phase one and two of the SOCOM framework have been integrated, phase three of the proposed framework is currently under development. In the near future, evaluation results of the SOCOM framework and suitability of matching algorithms will become available.

The SOCOM framework is semantic-oriented for two reasons. Firstly, during the ontology rendition phase, the context of an ontological resource is studied in order to determine the most appropriate translation result for its label. This context is defined by the semantics an ontology resource represents, which can be obtained by studying its surrounding concepts, i.e. super/sub-classes and property restrictions. Secondly, the mapping process makes uses of the similarity measures established in the ontology rendition phase in order to generate mapping results. The similarity measures are determined based on the semantics from each pair of ontology resources. An experimental version of the SOCOM framework has been integrated and is currently being evaluated.

Acknowledgement This research is partially supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation at Trinity College Dublin.

References

1. Powell S., Guru Interview: Sir Timothy Berners-Lee. KBE, October (2006)
2. Soergel D., Multilingual Thesauri and Ontologies in Cross-Language Retrieval. Presentation at the AAAI Spring Symposium on Cross-Language Text and Speech Retrieval (1997)

3. Simonet M., Diallo G., Multilingual Ontology Enrichment for Semantic Annotation and Retrival of Medical Information. MedNET, October (2006)
4. Shi C., Wang H., Research on Ontology-Driven Chinese-English Machine Translation. In Proceedings of NLP-KE, pp. 426-430 (2005)
5. Benjamins R. V., Contreras J., Corcho O., Gomez-Perez A., Six Challenges for the Semantic Web. SIGSEMIS Bulletin, April (2004)
6. Carpuat M., Ngai G., Fung P., Church W. K., Creating a Bilingual Ontology: A Corpus-Based Approach for Aligning WordNet and HowNet. In Proceedings of the 1st Global WordNet Conference (2002)
7. Malaise V., Isaac A., Gazendam L., Brugman H., Anchoring Dutch Cultural Heritage Thesauri to WordNet: Two Case Studies. In Proceedings of the Workshop on Language Technology for Cultural Heritage Data (2007)
8. Shimoji Y., Wada T., Dynamic Thesaurus Construction from English-Japanese Dictionary. In Proceedings of International Conference on Complex, Intelligent and Software Intensive Systems, pp. 918-923 (2008)
9. Euzenat J., Shvaiko P., Ontology Matching. Springer-Verlag (2007)
10. Pazienza M. T., Stellato A., An Open and Scalable Framework for Enriching Ontologies with Natrual Lanauge Content. In Proceedings of the 19th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (2006)
11. Pazienza M. T., Stellato A., Exploiting Linguistic Resources for Building Linguistically Motivated Ontologies in the Semantic Web. In Proceedings of the OntoLex workshop, Interfacing Ontologies and Lexical Resources for Semantic Web Technologies (2006)
12. Peters W., Montiel-Ponsoda E., Aguado de Cea G., Localizing Ontologies in OWL. In Proceedings of the OntoLex workshop, From Text to Knowledge: The Lexicon/Ontology Interface (2007)
13. Chang C., Lu W., The Translation of Agricultural Multilingual Thesaurus. In Proceedings of the 3rd Asian Conference for Information Technology in Agriculture, pp. 526-528 (2002)
14. Espinoza M., Gomez-Perez A., Mena E., LabelTranslator – A Tool to Automatically Localize an Ontology. In proceedings of the 5th European Semantic Web Conference, pp. 792-796 (2008)
15. Espinoza M., Gómez-Pérez A., Mena E., Enriching An Ontology with Multilingual Information. In Proceedings of the 5th European Semantic Web Conference, pp. 333-347 (2008)
16. Caracciolo C., Euzenat J., Hollink L., Ichise R., Issac A., Malaisé V., Meilicke C., Pane J., Shvaiko P., Stuckenschmidt H., Sváb-Zamazal O., Svátek V., Results of the Ontology Alignment Evaluation Initiative (2008)
17. Montiel-Ponsoda E., Peters W., Aguado de Cea G., Espinoza M., Gómez-Pérez A., Sini M., Multilingual and Localization Support for Ontologies. NeOn Deliverable (2008) Available at http://www.neon-project.org/web-content/images/Publications/neon_2008_ d242.pdf
18. Pazienza T. M., Stellato A., Linguistically Motivated Ontology Mapping for the Semantic Web. In Proceedings of the 2nd Italian Semantic Web Workshop, Semantic Web Applications and Perspectives (2005)
19. Trojahn C., Quaresma P., Bieira R., A Framework for Multilingual Ontology Mapping. In Proceedings of the 6th edition of the Language Resources and Evaluation Conference, pp. 1034-1037 (2008)
20. Trojahn C., Moraes M., Quaresma P., Vieira R., A Cooperative Approach for Composite Ontology Mapping. Journal on Data Semantics X, pp. 237-263 (2008)
21. Shvaiko P., Euzenat J., Ten Challenges for Ontology Matching. In Proceedings of the 7th International Conference on Ontologies, Databases and Applications of Semantics (2008)
22. O'Sullivan D., Wade, V., Lewis D., Understanding as We Roam. IEEE Internet Computing, 11 (2), pp.26--33 (2007)