Using Pseudo Feedback to Improve Cross-Lingual Ontology Mapping

Bo Fu, Rob Brennan, Declan O'Sullivan

Knowledge and Data Engineering Group, School of Computer Science and Statistics,
Trinity College Dublin, Ireland
`{bofu, rob.brennan, declan.osullivan}@cs.tcd.ie`

Abstract. Translation techniques are often employed by cross-lingual ontology mapping (CLOM) approaches to turn a cross-lingual mapping problem into a monolingual mapping problem which can then be solved by state of the art monolingual ontology matching tools. However in the process of doing so, noisy translations can compromise the quality of the matches generated by the subsequent monolingual matching techniques. In this paper, a novel approach to improve the quality of cross-lingual ontology mapping is presented and evaluated. The proposed approach adopts the pseudo feedback technique that is similar to the well understood relevance feedback mechanism used in the field of information retrieval. It is shown through the evaluation that pseudo feedback can improve the matching quality in a CLOM scenario.

Keywords: Cross-Lingual Ontology Mapping; Pseudo Feedback.

1   Introduction

One approach to ontology construction is to use language neutral identifiers to label concepts [1], whereby ontological entities are natural language independent. Given such ontologies, there would be little need for cross-lingual ontology mapping. However, as Bateman points out "the path towards viable ontologies is one that is irreconcilably connected to natural language" [2]. With this view to ontology construction being largely adopted in practice [3], multilinguality is increasingly evident in ontologies as experts with various natural language preferences build knowledge representations in multilingual organisations [4], government regulations [5], medical practice [6], to just name a few. As a result, notable research can be seen in the area of multilingual ontology acquisition [7], linguistic enrichment of ontologies [8] and ontology localisation [9]. These efforts highlight the importance of dealing with multilingual ontologies, and the ability to reason over knowledge bases regardless of the natural languages in them has become a pressing issue in digital content management. Ontology mapping techniques must be able to work with otherwise isolated ontologies that are labelled in diverse natural languages.

One way to achieve semantic interoperability across natural language barriers is by means of cross-lingual ontology mapping (CLOM). A valid approach to CLOM is to translate the labels of a given ontology to the natural language used by the other ontology(ies) first, and apply monolingual ontology matching techniques next, as

demonstrated in [10, 11, 12, 13]. A key challenge involved in this approach is ensuring the translated labels will maximise the final matching quality, since noisy translations could potentially pose negative impact on the monolingual matching tools as shown in [14]. Previous work [15] shows that selecting suitable translations is critical to the generation of quality CLOM results. Motivated by this requirement, this paper presents a novel approach that uses pseudo feedback, which is inspired by the well understood relevance feedback mechanism commonly used in information retrieval, to select ontology label translations as a way to improve CLOM. The proposed approach is evaluated against a baseline system in an experiment that uses the Ontology Alignment Evaluation Initiative (OAEI) 2009 benchmark dataset and the OAEI gold standard involving ontologies labelled in English and French. The evaluation results suggest that the pseudo feedback feature improves the CLOM quality comparing to the baseline system.

The remainder of this paper is organised as follows. Some related work is outlined in section 2. A pressing challenge for current CLOM approaches is discussed in section 3. To address this challenge, the pseudo feedback feature to improve CLOM is proposed in section 4. This proposed approach is evaluated in a CLOM experiment discussed in section 5. Finally, conclusions and future work are discussed in section 6.

2   Related Work

Current approaches to CLOM can be summarised as follows, manual processing [16], corpus-based [17], via linguistic enrichment [18], via indirect alignment [19] and translation-based [10, 11, 12]. An example of manual CLOM is discussed in [16], where an English thesaurus is mapped to a Chinese thesaurus by hand. Given large and complex ontologies, such a time-consuming and labour-intensive approach may be infeasible. Ngai et al. [17] use a bilingual corpus to align WordNet (in English) and HowNet (in Chinese), however, as such corpora are not always available to domain-specific ontologies, this approach may be unsuitable in some CLOM scenarios. Pazienza & Stellato [18] propose a linguistically motivated mapping approach and urge linguistically motivated ontology development, whereby ontologies would contain human-readable linguistic resources that can offer strong evidence in the mapping process. To facilitate this process, the OntoLing plug-in [20] was developed for the Protégé editor. However, as pointed out by the authors, this enrichment process is currently unstandardised. As a result, it can be difficult to build CLOM systems based upon such linguistically enriched ontologies. Jung et al. [19] demonstrate indirect alignment for multilingual ontologies in English, Korean and Swedish, given alignment A which is generated between ontology O_1 (i.e. in Korean) and O_2 (i.e. in English), and alignment A' which is generated between ontology O_2 and O_3 (i.e. in Swedish), mappings between O_1 and O_3 can be generated by reusing alignment A and A' since they both concern one common ontology O_2. Assuming the availability of A and A', this is an achievable approach. However, as this technique requires the very existence of A and A' which currently remains a challenge in itself, it can be difficult to apply this approach in some CLOM settings.

Translating ontology labels is a popular technique to convert a cross-lingual mapping problem into a monolingual mapping problem. Bilingual dictionaries, multi-lingual thesauri and off-the-shelf machine translation (MT) tools are often used as

media to bridge between different natural languages presented in the ontologies at hand. Zhang et al. [12] use a Japanese-English dictionary to translate the labels in the Japanese web directory into English first, before carrying out monolingual matching procedures using the RiMOM tool in the OAEI 2008 mldirectory[1] test case. Bouma [21] uses the multilingual EuroWordNet and the Dutch Wikipedia to align the GTAA thesaurus (in Dutch) to Wordnet and DBpedia (both in English). Wang et al. [10] use the GoogleTranslate service to translate digital library vocabularies before applying instance-based matching techniques to generate mappings among library subjects written in English, French and German. Trojahn et al. [13] incorporate the work presented in [14, 19] and uses the GoogleTranslate API as the translation medium to achieve CLOM. In addition, their tool is accompanied by a mapping reuse feature as proposed in [19]. The aforementioned research illustrates that translation can serve as a means to the completion of CLOM tasks, and MT may be sufficient to bridge between different natural languages in a given CLOM scenario, but just how suitable are these translations in the matching sense as opposed to the linguistic sense? This question is discussed in detail next.

3   The Challenge of Translation in CLOM

In the well studied field of MT, various techniques aiming to improve the quality of translation such as statistical MT, rule-based MT are designed, all equipped with the ability to disambiguate word senses. By nature, MT tools are intended to generate the most accurate translations in the linguistic sense, which is not necessarily a requirement in CLOM. This is because ontology matching techniques often rely on the discovery of lexical similarities as demonstrated in [14]. To achieve CLOM, translation is merely a stepping-stone to the actual goal which is generating correspondences between ontological entities. Consequently, translating source ontology labels is not centred around finding localised equivalents for them, but to select translations that can lead them to quality candidate matches in the target ontology. A translation may be accurate in the eyes of a linguist (i.e. linguistically correct), but it may not be appropriate (i.e. neglect matches) in the mapping context. In this paper, *an appropriate ontology label translation (AOLT) in the context of cross-lingual ontology mapping is one that is most likely to maximize the success of the subsequent monolingual ontology matching step*. This notion of AOLT in CLOM can be illustrated in the following example where the source ontology is in English and the target ontology is in French. A source concept *Ph.D. Student* has a candidate translation *Ph.D. Étudiant* which has a synonym *Étudiant au doctorat* (for example, by looking up from a thesaurus). The target ontology happens to have a class labelled *Étudiant au doctorat*, in this case, *Étudiant au doctorat* should be considered as the AOLT in this scenario since it is the terminology used by the target ontology and is most likely to lead to a mapping as a result.

Note that the work presented in this paper should not be confused with ontology localisation, whereby ontology labels are translated so that the given ontology is adapted "to a particular language and culture" [22]. In this paper, ontology labels are purposely translated so that the given ontologies can be best mapped. The AOLT

---

[1] http://oaei.ontologymatching.org/2008/mldirectory

process is concerned with searching for appropriate translations (from a mapping point of view) among a pool of candidate translations that are believed to be the ones most likely to enhance the matching ability of the subsequent monolingual matching step, but not necessarily the most linguistically correct translations (from a localisation point of view). Note this is not a natural language processing technique, the AOLT process does not attempt to disambiguate word senses.

4   Using Pseudo Feedback in CLOM

Ruthven & Lalmas [23] present an extensive survey on relevance feedback used in information retrieval (IR). Broadly speaking, there are three types of relevance feedback, explicit, implicit and blind (also known as pseudo feedback). Explicit feedback is obtained after a query is issued by the user and an initial set of documents is retrieved, the user marks these initial documents as relevant or not relevant, and the system retrieves a better list of documents based on this feedback by computing a single or multiple iterations. Implicit feedback works similarly but attempts to infer users' intentions based on observable behaviour. Pseudo feedback is generated when the system makes assumptions on the relevancy of the retrieved documents. Explicit user feedback in monolingual ontology matching and its effectiveness is successfully demonstrated by Duan et al. in [24], where the user marks the matches generated to be true or false. This paper expands on the feedback techniques that can be used in ontology mapping which is inspired by pseudo feedback in IR, it concerns a feedback mechanism without the involvement of a user in cross-lingual mapping scenarios.

When using feedback in the context of CLOM, an initial set of matches generated after the first iteration of the CLOM process can be thought of as the initial set of documents retrieved by an IR system, and the assumption made against document relevancy in IR becomes the process of assuming which candidate matches in the initial set are indeed correct. Similarity measures are often used to illustrate the confidence level of a matching tool in its conclusion of a matched entity pair, which can be used by pseudo feedback when making assumptions on correct matches. There are many types of similarity measures used in ontology matching as documented by Euzenat & Shvaiko [25]. Although currently there is no obvious method that is a clear success [26], similarity measures nonetheless are a way to perceive the probability of a match being correct or not. The CLOM approach presented in this paper incorporates pseudo feedback, whereby the system assumes after an initial execution that matches with confidence measures above a certain threshold are correct. It then examines how these matches are generated. Currently, this involves examining which translation media were used. The results of this examination then influence the selection of AOLTs in the second iteration of the system. An overview of this approach is presented in section 4.1, followed by its implementation in section 4.2.

4.1   Process Overview

Fig. 1 illustrates the CLOM process that integrates pseudo feedback. Given ontologies O_1 and O_2 that are labelled in different natural languages, CLOM is achieved in three main steps. Firstly, O_1 is transformed through the *ontology rendering* process as O_1', which has the same structure as O_1 but contains entities labelled in the natural

language that is used by O_2. Secondly, O_1' is matched to O_2 using *monolingual ontology matching* techniques to generate candidate matches. Thirdly, these matches are reviewed by the *match assessment* process, where assumptions are made to speculate on correctness. The pseudo feedback, containing the translation media used by these "correct" matches, is then processed by the AOLT selection in the second iteration of the CLOM system. Each of these steps is discussed next in detail.
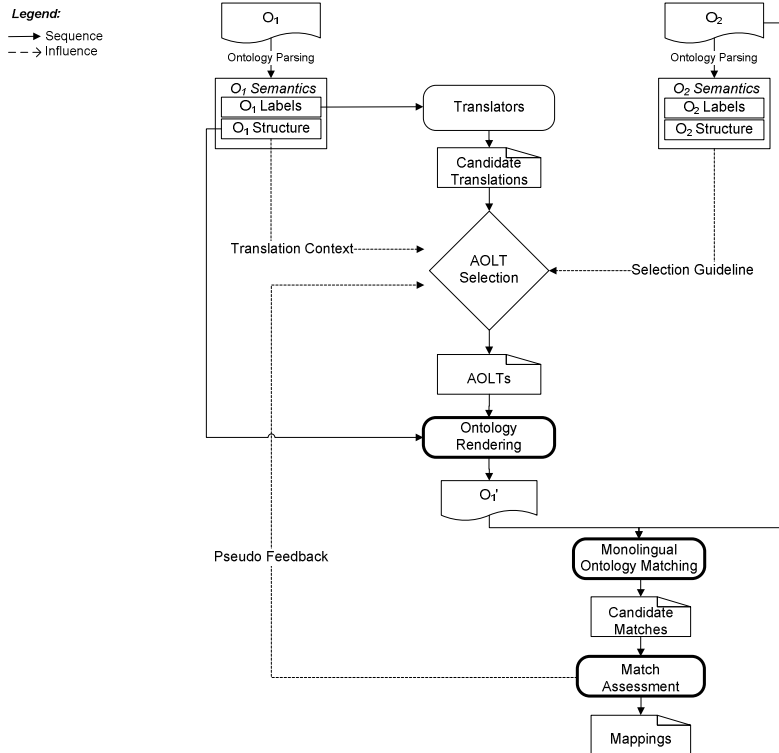

Fig. 1. Pseudo Feedback in CLOM

Ontology renditions are achieved by structuring the translated labels[2] in the same way as the original ontology, and assigning them with new base URIs to create well-formed ontology resources[3]. Zhao et al. [27] define ontology rendition as a process in the ontology development that consists of two roles, converting and interpreting. The converting role is the transformation of an ontology where the output has "formally different but theoretically equivalent" semantics, e.g. translating ontologies from

---

[2] In this paper, the translation of ontology labels refers to the translation of strings that are used to identify ontological resources in a formally defined ontology, e.g. the value of `rdf:ID` in `<Class rdf:ID="Thing"/>` or the fragment identifier, i.e. the string after the hash sign in `<owl:Class rdf:about="http://swrc.ontoware.org/ontology#Person"/>`. It does not refer to the content of `rdfs:label` elements such as `<rdfs:label>Thing</rdfs:label>`.

[3] The base URI is the unique identifier for an ontology and the resources within, as the resources in O_1' should not point to the original resources in O_1, new namespace declarations are assigned to the translated labels.

OWL to RDF via Web-PDDL [28]. The interpreting role renders formally specified commitments, which is the aim of the ontology rendering process shown in Fig. 1. Note that the structure of an input ontology is not changed during this process, as doing so would effectively alter the semantics of the original ontology.

The AOLT selection process makes use of the pre-defined semantics in the given ontologies and is concerned with identifying the most appropriate translations for a specific mapping scenario. To achieve this, firstly, for each extracted label in O_1, it is sent to the *translators* to generate candidate translations. Secondly, to identify the AOLTs in the specified CLOM scenario for an ontology label, the selection process takes the following into account:

- The *semantics in O_1* 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 can be collected to represent the context of use. For example, for a class node, its context can be represented by the labels of its super/sub/sibling-classes. For a property node, its context can be represented by the labels of the resources which this property restricts. For an individual of a class, its context can be characterised by the label of the class it belongs to.

- As O_1' is rendered so that its representation of O_1 can be best mapped to O_2, the *semantics in O_2* therefore act as broad AOLT selection guidelines. For example, when several translation candidates are available for a label in O_1, the most appropriate translation is the one that is most similar to what is used in O_2, e.g. the example given in section 3.

Once AOLTs are identified, O_1'[4] is generated and various monolingual matching techniques can be applied to create correspondences between O_1' and O_2. These matches are finally sent to the *match assessment* process, and "correct" matches are assumed to be those that have confidence measures above a specified threshold. Based on this assumption, pseudo feedback is generated which contains the most effective translation media for the particular ontologies at hand. In the second iteration of the CLOM system, the translations returned from these media are perceived to be the AOLTs. This process is further demonstrated and explained with an example in section 5.1.

4.2   Implementation

An implementation of the proposed approach is shown in Fig. 2. The Jena Framework[5] 2.5.5 is used to parse the ontologies, extract entity labels and to generate surrounding resource labels for a given entity. Candidate translations of the source ontology labels are obtained from the *machine translation service* that uses the GoogleTranslate[6] API 0.5 and the WindowsLive[7] translator. These candidate translations are stored in the *translation repository* and formatted in XML. For the ontology pair shown in Fig. 3, Fig 3a presents a snippet of the translation repository generated for the source ontology labelled in English, and Fig. 3b shows a snippet of

---

[4] O_1' exists only for purpose of the mapping, it should not be considered as a localised O_1.

[5] http://jena.sourceforge.net

[6] http://code.google.com/p/google-api-translate-java

[7] http://www.windowslivetranslator.com/Default.aspx Note at the time of writing, the Windows Live translator has been renamed as the Bing translator.

the lexicon repository generated for the target ontology labelled in French. Ontology labels are often concatenated (as white spaces are not allowed in the OWL/RDF naming conversion), which cannot be processed by the integrated MT tools. To overcome this issue, concatenated ontology labels (stored in the *OntLabel* attribute in Fig. 3) are first split into sequences of their constituent words (as machine readable values and stored in the *MRLabel* attribute in Fig. 3) before passed to the MT tools. In the example shown in Fig. 3, as capital letters are used to indicate the beginning of another word, white spaces are inserted before each capital letter found other than the first one. A *lexicon repository* is generated that contains the target ontology labels, their corresponding synonyms and surroundings. An example of this is shown in Fig. 3b. Synonyms are generated by calling the *lexicon dictionary service*, which queries synonyms-fr.com[8] for synonyms in French. The generations of the translation repository and the lexicon repository take place in parallel. Finally, both repositories are stored in the eXist DB[9] 1.0rc.
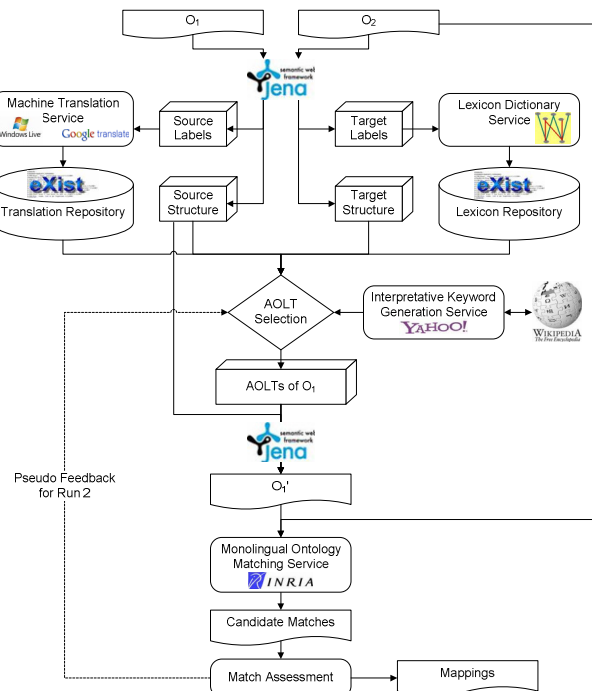

Fig. 2. An Implementation of Pseudo Feedback in CLOM

The AOLT selection process queries the repositories to compare the candidate translations of an ontology label to the resources stored in the lexicon repository. If matches (to a target label or synonym of a target label) are found, preference is always given to what is used by the target ontology (e.g. when a candidate translation is linked to a target label's synonym, this synonym's corresponding label that appears in the target ontology is deemed to be the AOLT). If no match is found, for each

---

[8] http://synonyms-fr.com
[9] http://exist.sourceforge.net

candidate, a set of interpretative keywords are generated to illustrate the meaning of this candidate. This is achieved by querying Wikipedia[10] via the Yahoo Term Extraction Tool[11]. Using a space/case-insensitive edit distance string comparison algorithm based on Nerbonne et al.'s method [29], the candidate with keywords that are most similar to the source label's semantic surrounding is chosen as the AOLT. Finally, AOLTs are concatenated to construct well-formed resource labels (stored as values in the attribute *ConLabel* of the *Candidate* element in Fig. 3a and the *Syn* element in Fig. 3b) by replacing white spaces with underscores. Once the AOLTs are determined for source labels, given the original ontology structure, O_1' is generated using the Jena framework, and matched to O_2 using the Alignment API[12] 3.6.

```
              (Source Ontology)              (Target Ontology)

Publication                       Publication
    ├── Article                       ├── Livre
    └── Report                        ├── Actes
            ├── ProjectReport         └── Rapport
            └── TechnicalReport
```

```
…
<Resource id="CLS-1" OntLabel="Article" MRLabel="Article"/>
    <Translation>
        <Candidate id="CDD-0" value="L'article" source="google" ConLabel="L'article"/>
        <Candidate id="CDD-1" value="Article" source="wl" ConLabel="Article"/>
    </Translation>
    <Surrounding id="CLS-0" OntLabel="Publication"/>
    <Surrounding id="CLS-2" OntLabel="Report"/>
</Resource>
…
<Resource id="CLS-3" OntLabel="ProjectReport" MRLabel="Project Report"/>
    <Translation>
        <Candidate id="CDD-7" value="Rapport de projet" source="google"
        ConLabel="Rapport_de_projet"/>
        <Candidate id="CDD-8" value="Rapport de projet" source="wl"
        ConLabel="Rapport_de_projet"/>
    </Translation>
    <Surrounding id="CLS-2" OntLabel="Report"/>
    <Surrounding id="CLS-4" OntLabel="TechnicalReport"/>
</Resource>
…
```

<center>(a) Translation Repository – An Example</center>

```
…
<Resource id="CLS-0" OntLabel="Publication" MRLabel="Publication"/>
    <Synonym>
        <Syn id="SYN-0" value="Parution" source="synonyms-fr.com" ConLabel="Parution"/>
        <Syn id="SYN-1" value="Sortie" source="synonyms-fr.com" ConLabel="Sortie"/>
        <Syn id="SYN-2" value="Ouvrage" source="synonyms-fr.com" ConLabel="Ouvrage"/>
    </Synonym>
    <Surrounding id="CLS-1" OntLabel="Livre"/>
    <Surrounding id="CLS-2" OntLabel="Actes"/>
    <Surrounding id="CLS-3" OntLabel="Rapport"/>
</Resource>
…
```

<center>(b) Lexicon Repository – An Example</center>
<center>**Fig. 3.** Examples of the Translation Repository & the Lexicon Repository</center>

Collisions can occur when more than one entity in O_1 concludes with the same value as its AOLT. A summary of collision solutions is presented in Table 1. When a collision is detected between two entities, priority is given to the one that was influenced by the target ontology (e.g. derived based on a match to target label or synonym) as scenario *i*, *ii*, *iii* and *iv* illustrate in Table 1. If both entities arrive to the same AOLT with an equal strategy (e.g. both came from a match made to a target

---

label's synonym) as shown in Table 1 scenario *v*, *vi* and *vii*, the later entity will seek the next translation in line. If no more alternative translation is available to this later entity, a numerical number (that is checked to be free of collision) is attached to the collided term. This ensures that both entities will have well-formed (i.e. unique) URIs. These numbers are selected at random with the intent of avoiding the introduction of any kind of patterns into the translation selection process.

Table 1. Resolving Translation Collision

Collision Scenario	AOLT Selection Strategy		Solution
	Entity 1	Entity 2	
i	candidate translation matches target label's synonym	candidate translation matches target label	entity 2 keeps the collided AOLT; entity 1 seeks alternative translation
ii	derived from interpretative keyword comparison	candidate translation matches target label's synonym	
iii	candidate translation matches target label	candidate translation matches target label's synonym	entity 1 keeps the collided AOLT; entity 2 seeks alternative translation
iv	candidate translation matches target label's synonym	derived from interpretative keyword comparison	
v	candidate translation matches target label	candidate translation matches target label	
vi	candidate translation matches target label's synonym	candidate translation matches target label's synonym	
vii	derived from interpretative keyword comparison	derived from interpretative keyword comparison	

Given the matching results[13] generated by the Alignment API and the origins of all AOLTs, the *match assessment* process assumes that matches with at least 0.5 confidence levels are correct[14] and computes a set of statistical feedback based on this assumption. This feedback contains the usage (as percentages) of each translation medium used by the "correct" matches. The translations which are generated by the highest ranked (i.e. highest usage) MT tools are prioritised in the AOLT selection process during the second execution of the system. This is further illustrated with an example in section 5.1.

5   Evaluation

To evaluate the effectiveness of the proposed pseudo feedback mechanism, the implemented system is compared to a baseline system that is already proven effective in [15]. The only difference between the two approaches is the pseudo feedback. The baseline system integrates the same set of tools and APIs, except that it does not attempt to use pseudo feedback to influence the selection of AOLTs (i.e. the baseline system is the first iteration of the implementation discussed in section 4.2). The evaluation experiment uses the OAEI 2009 benchmark dataset involving ontologies of

---

[13] In the Alignment API, a match between a source ontology resource and a target ontology resource is represented with a relation and accompanied by a confidence level that range between 0 (not confident) and 1 (confident).

[14] As confidence levels range between 0 and 1, 0.5 is a natural division point where matches would either incline towards being either confident (i.e. equal or above 0.5) or not confident (i.e. below 0.5). This threshold on confidence measure cannot be configured by the user in the current implementation, as this paper is a proof of concept of whether pseudo feedback can be applied in CLOM rather than looking for a best feedback configuration in CLOM. The pseudo feedback presented in this paper speculates on which matches could be correct, it is not designed as an accurate assessment of the matches generated.

the bibliography domain labelled in English and French[15]. Its setup is discussed in section 5.1, followed by its findings in section 5.2.

5.1   Experimental Setup

Fig. 4 illustrates an overview of the experimental setup. Ontology 101 is labelled in English and has 36 classes, 24 object properties, 46 data type properties and 137 instances. Ontology 206 contains similar semantics, except it has one less object property and is labelled in French. The English ontology is matched to the French ontology using the proposed approach with pseudo feedback and the baseline approach to generate mappings M_F and M_B respectively, using eight matching algorithms[16] that are supported by the Alignment API.
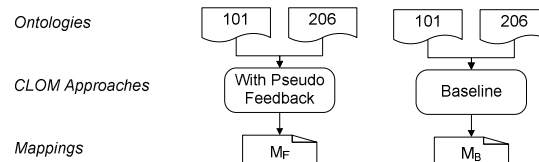
Fig. 4. Experiment Overview

Note that the original OAEI test scenario does not involve any translations of ontology labels. It was designed to examine the strength of structure-based monolingual matching techniques since ontology 101 and 206 have highly similar structures. Though this is not the goal of the evaluation setup presented in this paper, nevertheless, this test case provides us with a pair of ontologies in different natural languages and a reliable gold standard[17] for the evaluation of M_F and M_B.

For each matching algorithm executed, the pseudo feedback mechanism selects the matches with at least 0.5 confidence levels, and investigates how the AOLTs were determined among these "correct" matches. For example, the pseudo feedback generated after the first iteration of the system when using the StrucSubDist-Alignment matching algorithm is shown in Fig. 5. The attributes of the root element include the matching algorithm used (stored in the `algorithm` attribute in Fig. 5), the cut-off point of the assumption (stored in the `threshold` attribute in Fig. 5), the total matches generated by the specified matching algorithm (stored in the `matches` attribute in Fig. 5) and the assumed-to-be correct matches found (stored in the `estimate` attribute in Fig. 5). In the case for the StrucSubsDistAlignment matching algorithm shown in Fig. 5, at a threshold of 0.5, a total of 86 correct matches are identified within a set of 103 matches. Each `<Entry>` element records the total count (stored in the `count` attribute in Fig. 5) of a particular translation medium used (stored in the `medium` attribute in Fig. 5) and its accumulated usage (stored in the `usage` attribute in Fig. 5, which is calculated as *count/estimate*). In Fig. 5, the pseudo feedback indicates that firstly, the majority of AOLTs originated from the target ontology (i.e. either labels used in the target ontology or synonyms of these labels).

Secondly, it shows that the same translations were returned by the integrated MT tools at times. In such cases, it would not be fair to credit either MT tool, it is therefore categorised on its own (ranked second highest in the example shown in Fig. 5). Thirdly, a greater number of AOLTs came from the GoogleTranslate API (in third rank) than the WindowsLive translator (in fourth rank) when using the StrucSubsDist-Alignment algorithm in this particular experiment. Finally, it shows that a small number of matches are made between externally defined resources (e.g. `rdf:resource ='http://www.w3.org/1999/02/22-rdf-syntax-ns#List'` as defined by the World Wide Web Consortium) which are categorised in fifth.

```
<PseudoFeedback algorithm="StrucSubsDistAlignment" threshold="0.5" matches= "103.0"
estimate="86.0" >
    <Entry count="31.0" medium="TargetOntology" usage="0.360465"/>
    <Entry count="23.0" medium="BothMT" usage="0.267441"/>
    <Entry count="17.0" medium="Google" usage="0.197674"/>
    <Entry count="12.0" medium="WindowsLive" usage="0.139534"/>
    <Entry count="3.0" medium="External" usage="0.034883"/>
</PseudoFeedback>
```

Fig. 5. An Example of Pseudo Feedback

In the second iteration of the system using the StrucSubsDistAlignment algorithm, the strategy for which translation media to use is thus determined by the order shown in Fig. 5. For a source ontology label, when a translation that originates from the target ontology is available, it is chosen as the AOLT; if not, use the translation that is agreed by both MT tools; in the absence of these two options, choose the translation returned from the GoogleTranslate API; if all fails, use the translation returned from the WindowsLive translator. This feedback to the AOLT selection process is repeated for all other matching algorithms in the second run of the system. Note that the ranking of the translation media is not necessarily always the same with what is shown in Fig. 5, as it depends on the statistics generated by the pseudo feedback which varies by the matching techniques used.

In addition to what is discussed in Table 1, new rules are included in the collision resolution process for the second iteration of the system. Priority is given to higher ranked MT media. For example, when two entities both choose the same value as its AOLT, the system checks how they each arrived to this conclusion. The higher ranked translation strategy will keep the collided term as its AOLT, and the other entity will seek for an alternative from a lower ranked MT medium. It is possible that a collision is unsolved still when all other alternatives cause further collisions or simply do not exist. In such cases, for the entity that is seeking an alternative translation, a unique numerical number is attached to the end of collided term as explained previously in section 4.2.

5.2 Experimental Results

M_F and M_B were evaluated based on the gold standard (see footnote 15) provided by the OAEI. Firstly, the evaluation identifies the correct matches in M_F and M_B based on the gold standard, computes and compares their respective precision, recall and F-measure scores. A correct mapping is one that is included in the gold standard regardless of its confidence measure. Precision (shown in Fig. 6a), recall (shown in Fig. 6b) and F-measure (shown in Fig. 6c) scores were calculated for all eight experimented matching algorithms. Fig. 6 shows higher precision, recall and F-measure scores for the matches found in M_F across all matching algorithms. On

average, M_B has a precision of 0.7355, a recall of 0.5928 and an F-measure of 0.6428, which have all been improved when the pseudo feedback mechanism is incorporated, leading to an average precision of 0.7875, recall of 0.6268 and F-measure of 0.6873 in M_F. These statistics indicate that M_F not only contains a greater number of correct matches, but also is more complete than M_B.
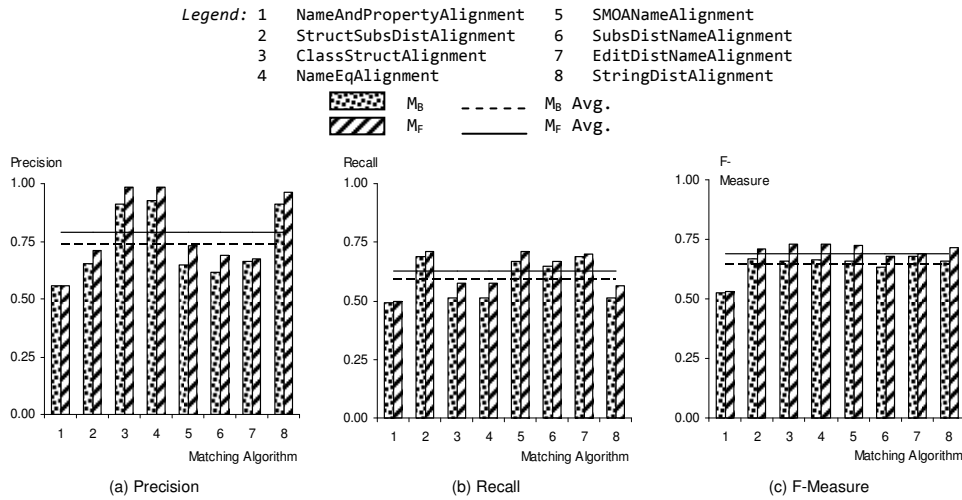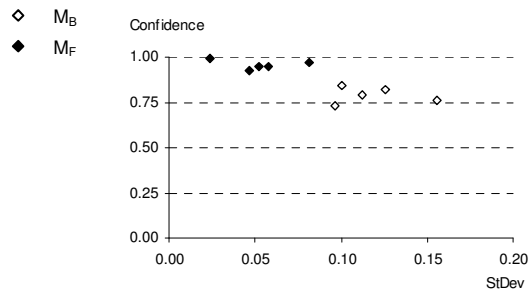

Fig. 6. Precision, Recall and F-Measure Overview

Fig. 7. Confidence Measures in M_B and M_F

Secondly, as confidence levels are not accounted by precision, recall or F-measure, for the correct matches found in M_B and M_F, their confidence means and standard deviations were also calculated. The mean is the average confidence of the correct matches found in a set of matches, where higher means indicate more confident results. The standard deviation is a measure of dispersion, where the greater it is, the bigger the spread in the confidence levels. Higher quality matches therefore are those with high confidence means and low standard deviations. Note, some matching algorithms (e.g. NameEqAlignment, StringDistAlignment and ClassStruct-Alignment which incorporated StringDistAlignment in the experiment) only created matches with 1.0 confidence levels, therefore were not included in this study. Fig. 7 presents an overview on the evaluation of the confidence means and standard deviations. The correct matches in M_F are always higher in confidence and lower in dispersion. This finding indicates that in addition to improving the precision, recall

and F-measure, the pseudo feedback feature can also facilitate monolingual matching techniques in their ability to generate correct matches more confidently.

It may be argued that as the differences shown in the F-measure scores between M_B and M_F are relatively small, it would be difficult to conclude an improvement in M_F. To validate the statistical significance of the findings so far, and to validate the difference (if it exists) between the two approaches, paired t-tests were carried out on the F-measure scores across all matching algorithms and a p-value of 0.003 is found. At a significance level of $\alpha=0.05$, this p-value rejects the null hypothesis (null hypothesis being there is no difference between the two CLOM approaches) and indicates that the findings are statistically significant. This further confirms the effectiveness of the pseudo feedback mechanism in the experiment.

It should be noted that the experimental setup is somewhat limited in the size of the ontologies used, their comparable natural languages and structures (as discussed in section 5.1). Nonetheless, the evaluation results from this experiment do suggest a positive impact of pseudo feedback and its ability to facilitate monolingual ontology matching techniques in the process of generating quality CLOM results.

6   Conclusions and Future Work

This paper presents a novel approach to CLOM that incorporates the pseudo feedback technique that is similar to the well-established relevance feedback mechanism used in the field of information retrieval. The proposed approach makes assumption on the correct matches in an initial matching set that is generated after the first iteration of the CLOM system. The pseudo feedback mechanism then determines how these "correct" matches are generated by detecting the translation media used, and finally sends this feedback back to the system to aid the selection of ontology label translations in the second iteration of the CLOM system. The advantages of the proposed approach are demonstrated using an OAEI dataset and evaluated against the OAEI gold standard. Based on the experimental findings presented in this paper, there are indications that the proposed pseudo feedback feature enhances the performance of the monolingual ontology matching techniques used.

Several potential future research directions can be derived from the findings presented in this paper. Firstly, the use of feedback in CLOM can be expanded to incorporate explicit and implicit feedback, whereby user knowledge and user behaviours may be used to assist the generation of reliable mappings. Secondly, current implementation of the proposed approach in this paper can be extended. For example, the current pseudo feedback mechanism assumes that correct matches are above the 0.5 confidence level, future implementations may include several thresholds that can be configured by users. The pseudo feedback can also be further extended to implement negative feedback (i.e. a blacklist as opposed to a whitelist of translation media as shown in this paper) so that the AOLT selection process recognises what not to do in a given mapping scenario. Additionally, MT tools in the current implementation are not specialised to work with highly refined domains such as medical ontologies. This may be improved given domain-specific translation tools. Thirdly, the risks involved and their impact (e.g. when the assumptions made on the "correct" matches are simply invalid) on the CLOM quality when applying pseudo feedback in CLOM is not yet investigated in this paper, future research could explore

this area. Fourthly, only two iterations of the CLOM system is demonstrated in this paper, further iterations of the system using pseudo feedback can be evaluated in order to investigate whether a third, fourth etc. iteration of the process can further improve mapping quality. Lastly, the ontologies used in the experiment shown in this paper are relatively small in size of the same domain, with comparatively similar natural language pairs and structures, the scalability and the effectiveness of the proposed approach should be tested against large ontology pairs with overlapping domains that involve more distinct natural languages and structures. This is currently being investigated as part of the on-going research.

References

1. Nirenburg S., Raskin V., Ontological Semantics, Formal Ontology, and Ambiguity. In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems, 151-161 (2001)
2. Bateman J. A., Ontology Construction and Natural Language. Workshop on Formal Ontology in Conceptual Analysis and Knowledge Representation (1993)
3. Noy N. F., McGuinness D. L., Ontology development 101: A guide to creating your first ontology. Technical Report SMI-2001-0880, Stanford Medical Informatics (2001)
4. Caracciolo C., Sini M., Keizer J., Requirements for the Treatment of Multilinguality in Ontologies within FAO. In Proceedings of the 3rd International Workshop on OWL: Experiences and Directions (2007)
5. Kerremans K., Temmerman R., Tummers J., Representing Multilingual and Culture-Specific Knowledge in a VAT Regulatory Ontology: Support from the Termontography Method. In Proceedings of the OTM Confederated International Workshops, LNCS 2889, 662-674 (2003)
6. Chen H., Wang Y., Wang H., Mao Y., Tang J., Zhou C., Yin A., Wu Z., Towards a Semantic Web of Relational Databases: a Practical Semantic Toolkit and an In-Use Case from Traditional Chinese Medicine. In Proceedings of the 5th International Semantic Web Conference, LNCS 4273, 750-763 (2006)
7. Nichols E., Bond F., Tanaka T., Fujita S., Flickinger D., Multilingual Ontology Acquisition from Multiple MRDs. In Proceedings of the 2nd Workshop on Ontology Learning and Population, 10-17 (2006)
8. Buitelaar P., Cimiano P., Haase P., Sintek M., Towards Linguistically Grounded Ontologies. In Proceedings of the 6th European Semantic Web Conference, LNCS 5554, 111-125 (2009)
9. Peters W., Montiel-Ponsoda E., Aguado de Cea G., Localizing Ontologies in OWL. In Proceedings of the OntoLex 2007 Workshop (2007)
10. Wang S., Isaac A., Schopman B., Schlobach S., Van der Meij L., Matching Multi-lingual Subject Vocabularies. In Proceedings of the 13th European Conference on Digital Libraries, LNCS 5714, 125-137 (2009)
11. Trojahn C., Quaresma P., Vieira R., A Framework for Multilingual Ontology Mapping. In Proceedings of the 6th edition of the Language Resources and Evaluation Conference, 1034-1037 (2008)
12. Zhang X., Zhong Q., Li J., Tang J., Xie G., Li H., RiMOM Results for OAEI 2008. In Proceedings of the 3rd International Workshop on Ontology Matching, 182-189 (2008)

13. Trojahn C., Quaresma P., Vieira R., An API for Multi-lingual Ontology Matching. In Proceedings of the 7th Conference on International Language Resources and Evaluation, ISBN 2-9517408-6-7, 3830-3835 (2010)
14. Fu B., Brennan R., O'Sullivan D., Cross-Lingual Ontology Mapping – An Investigation of the Impact of Machine Translation. In Proceedings of the 4th Asian Semantic Web Conference, LNCS 5926, 1-15 (2009)
15. Fu B., Brennan R., O'Sullivan D., Cross-Lingual Ontology Mapping and Its Use on the Multilingual Semantic Web. In Proceedings of the 1st Workshop on the Multilingual Semantic Web, CEUR Vol. 571, 13-20 (2010)
16. Liang A., Sini M., Mapping AGROVOC & the Chinese Agricultural Thesaurus: Definitions, Tools Procedures. New Review of Hypermedia & Multimedia 12(1), 51-62 (2006)
17. 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)
18. Pazienta M., Stellato A., Linguistically Motivated Ontology Mapping for the Semantic Web. In Proceedings of the 2nd Italian Semantic Web Workshop, 14-16 (2005)
19. Jung J. J., Håkansson A., Hartung R., Indirect Alignment between Multilingual Ontologies: A Case Study of Korean and Swedish Ontologies. In Proceedings of the 3rd International KES Symposium on Agents and Multi-agent Systems – Technologies and Applications, LNAI 5559, 233-241 (2009)
20. Pazienza M. T., Stellato A., Exploiting Linguistic Resources for Building Linguistically Motivated Ontologies in the Semantic Web. In Proceedings of OntoLex Workshop 2006: Interfacing Ontologies and Lexical Resources for Semantic Web Technologies (2006)
21. Bouma G., Cross-lingual Ontology Alignment using EuroWordNet and Wikipedia. In Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC'10), ISBN 2-9517408-6-7, 1023-1028 (2010)
22. 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 (2008)
23. Ruthven I., Lalmas M., A survey on the use of relevance feedback for information access systems. Knowledge Engineering Review 18, 2, 95-145 (June 2003)
24. Duan S., Fokoue A., Srinivas K., One size does not fit all: Customizing Ontology Alignment using User Feedback. In Proceedings of the 9th International Semantic Web Conference, LNCS 6497 (2010)
25. Euzenat J., Shvaiko P., Ontology Matching. Springer-Verlag Berlin/Heidelberg (2007)
26. Ichise R., Evaluation of Similarity Measures for Ontology Matching. In Proceedings of the 22nd Annual Conference of the Japanese Society for Artificial Intelligence, LNAI 5447, 15-25 (2009)
27. Zhao G., Zheng J., Meersman R., An Architecture Framework for Ontology Development. In Proceedings of the IADIS International Conference (2003)
28. Dou D., McDermott D., Qi P., Ontology Translation on the Semantic Web. Journal on Data Semantics II, 35-57 (2004)
29. Nerbonne J., Heeringa W., Kleiweg P., Edit Distance and Dialect Proximity. In: Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison, 2nd ed. CSLI, Stanford, v-xv. (1999)