Investigating Core Technologies in Computer-aided Multi-lingual Translation Memory

Huangfu Wei
School of Foreign Languages, North China Electric Power University, Beijing 102206, China
mailtohf@163.com

Abstract - Based on status quo analysis of the current machine translation and computer-aided translation, this study intends to look into the core technologies of multilingual translation memory. This study takes the holistic perspective that views example-based machine translation and computer-aided translation memory as two contrasting and also complementary solutions to account for the complexity and diversity and world languages in translation. Also, Technological solutions to a multi-lingual translation project are discussed in accordance with the problem analyses. Three core technologies are brought into detailed analysis. In calculating sentence similarities, this study argues in favor of a mixed method of minimum edit distance. In structuring a multi-lingual translation memory, an exemplary structure in XML format is proposed. And for translation cleaning up, the routine practice is adjusted to fit a CAT system with a multi-lingual translation memory. This paper ends with a concluding remark of this study and by bringing up the future work.

Index Terms - multi-lingual translation memory, sentence similarity, TM structure, translation cleaning up.

I. Introduction

With only 70% of readability and 20% of accuracy(Dorr, Bonnie, 2010), machine translation (MT) fails to satisfy the increasing demand for high quality translated texts, while computer-aided translation is gaining recognition from all sides. Putting this trend in perspective, it can be said that it is through human-machine interaction that higher quality and efficiency in translation can be achieved. In some specialties, such as computer, telecommunication, automation and military, where the repetition rate can reach as high as 20%-70%, computer-aided translation memory (TM) can cut down on large amount of repetitive work.

Computer-aided translation (CAT) has a long history abroad and has been fully developed by a number of companies with such products as TRADOS, SDLX, Dejavu, Star, Al2tavista, Transit, TransSuite2000, EurolangOptimize, IBM Translation Manager, WordFisher, Wordfast, OmegaT. CAT is well accepted and used in western countries by translators and technicians in translating science and technology texts(Lagoudaki, 2006). But in China with Yaxin and Xueren and Transmate CATS supporting only English-Chinese bilingual translation, this technology is still held in contempt and underestimated by technicians and little known and used by translators.

Considering the fast development of CAT technology and increasing demand for multi-lingual CAT tools, this paper is to investigate the key technologies in computer-aided multi-lingual translation memory and their corresponding algorithm so as to shed some light on future multi-lingual translation memory platforms.

II. EBMT and Computer-aided TM

A. Example-based Machine Translation

Makoto Nagao(1984) first proposed the EBMT method from the perspective that translation is not done by deep structure analysis of the sentences, rather the sentences are segmented into translation units, such as clauses and phrases, and then these segments are restructured and translated into target language. Based on this theory, machine translation starts with calculating the source sentences' similarities with those in a bilingual corpus and find out the closest ones as references for direct uses or improvement. Therefore, the translation is generated from the examples in the corpus, which are more reliable and accurate. If the corpus size is large enough, the input source sentences will be translated into the target language with high quality. Compared with the rule-based machine translation, which relies on manmade rules and can not deal with exceptions and has contradictory rules in need of manual debugging, the EMBT method is at an advantage (Carl, 2005).

But, the EMBT method needs a considerably large reference corpus with bilingual sentence examples, which need huge human and financial resources to build. If the coverage and size of the corpus is limited, computers cannot find perfect matches for the translation. Moreover, the sentence similarity calculation is the bottleneck of this method because none of the available method can account for all the differences between languages(Ramiz,2009;Palakorn,2008). It is for these reasons that the EMBT method cannot be used to replace human translation and only as a plug-in method to improve efficiency and quality of transition.

B. Computer-aided Translation Memory

Translation memory (TM) is the core part of any CAT system. Bowker defines TM as the database to store source and target language sentence pairs. CAT tools automatically use database to provide translators with referential sentences. The database can enlarge itself by receiving more sentences from translators in translation new sentences or editing older translation, or by aligning bilingual files and sending them to

This work is supported by "the Fundamental Research Funds for the Central Universities" in China #13MS47

TM. The bilingual corpus is usually the basis of TM and its size and coverage is enlarged through human-machine interface.

Translation memory is especially helpful if the documents have a certain amount of repetition, otherwise its use will be limited. Moreover, the translation memory is empty and need expanding to be more effective. However, in reality, the needed repetition of documents is not always met, and the size and coverage of the TM cannot be large enough to always provide high-quality translation due to languages' infinite ability to generate sentences. Also, there is no linguistic annotation for the data in TM, and deep mining of the data are needed for further application.

III. Core Technologies in Computer-aided TM

In the work of multi-lingual translation, computer-aided translation memory technologies need to be constantly improved to have optimal use of both human and computers. In this process, key technologies may include the following ones: designing a multi-lingual translation memory, and calculating the similarities between sentences in source languages and those in translation memory and providing users with the most similar references, and using the references to pseudo-translate the source languages into target languages.

A. Calculating Sentence Similarities

In building translation memory, deciding similarities between translation units in source sentences and target reference sentences is the key to the efficiency and quality of the translation. This procedure can also be taken as an analogical mapping, in which the translation units in both the source and target are compared and analogical translation results are generated from translation memory (Sandipan, 2012). Four possible approaches to calculation of similarities are first summarized as follows:

1) String-based Algorithm: There have been many different approaches to a string-based algorithm, among which the edit distance is the most common and simplest, also called Levenshtein distance. Let the length of the word P and W is n and m respectively, then the edit distance, i.e. ed(P, W), is the calculation of matrix of n rows and m columns. Then, edit(i,j)=min(edit(i-1,j)+1,edit(i,j-1)+1,edit(i-1,j-)+ed(Pi,Wj)), among which Pi is the ith character in word P and Wj is the jth character in word W. If Pi=Wj, then ed(Pi,Wj)=0, else ed(Pi,Wj)=1. The value at n row and m column is the edit distance.

2) VSM-based Algorithm: With vector space model (VSM), the calculation of similarities of character sequences a and b can be made with cosine of VSM values v1 and v2 of characters s1 and s2. The smaller of the vector space angles of these two characters are, the greater their similarities are. The formula is as follows:

$$sim(s1,s2)=\cos(v1,v2)=\frac{\sum_{m=0}^{n} c_{1m} * c_{2m}}{\sqrt{\sum_{m=0}^{n} c_{1m}{}^{2}}\sqrt{\sum_{m=0}^{n} c_{2m}{}^{2}}}$$

3) Word-based Agorithm: There are two types of word-based algorithms, i.e. algorithm of ontology based on thesauruses or semantic dictionary and that based on large corpus. In the corpus-based algorithm, the frequency distribution of the target words' contexts is used in calculating their similarities, which needs a large corpus and great amount of probability calculation. Another method calculates the distance between the nodes of words in a semantic dictionary, such as Hownet, WordNet, TongYiCi Cilin, etc. In terms of both Chinese and English word senses, sememes used in Hownet represent word senses and thus the bases of word similarities are sememes' similarities, which are converted from their semantic distance by the following formula with A and B representing the sememes, d representing their distance between their nodes, α representing an adjustable parameter:

$$Sim(A,B)=\frac{\alpha}{d+\alpha}$$

4) Syntax-based Algorithm: This method relies on the deep structure analysis of the target reference sentences and source sentences and uses parsers to generate the dependency tree, which will be used in calculating sentence similarities. To reduce the noises and increase accuracy, calculations are only made in the effective collocation pairs and the value of a weight of every pair, i.e. keywords and their dependent effective words. The formula is as follows:

$$Sim(Sen1,Sen2)=\frac{\sum_{i-1}^{n} Wi}{Max\{Pair1,Pair2\}}$$

All in all, for a multi-lingual translation project, distinctive features of every language bring about difficulties in similarity calculation. Considering the advantages of the above mentioned methods, a minimum edit distance method may be more desirable to deal with this issue. From the Levenshtein distance algorithm, the number of words in need of matching in target and source sentences is first decided, and then their similarities can be calculated with the following formula:

$$Sim(A,B)=\frac{Max(A_{Length},B_{Length})-d[n,m]}{Max(A_{Length},B_{Length})}*100\%$$

The procedure to get d[n,m] is as follows: firstly, calculate length of the source sentence n and that of the target sentence m; secondly, initiate a matrix d of n+1 rows and m+1 columns; thirdly, let s[i] the i^{th} in the source sentence and t[j] the j^{th} word in the target sentence, then if s[i]= t[j], source=0, else source=1. Thus, d[n,m] is the minimum value of d[i-1,j]+source, d[i,j-1]+source and d[i-1,j-1]+source. The Sim(A,B) is in terms of percentage in the range of 0%-100% of similarities between sentence A and B. It is so that a threshold value of 75% should also be preset as default or adjusted according to requirement.

B. Structuring a Multi-lingual Translation Memory

As the core part of any CAT tool, translation memory provides the referential sentences and the basis of target-language generation or translation cleaning up. The overall design includes the translation unit storage in the TM and managing TM database. Generally speaking, the TM database stores the translation unit at a sentential level, phrasal level or word level alignment and deliver the useful information in right granularity and form to the translators. Considering the convenience and ease of data management and expansion, a sentential level alignment of translation units is at an advantage in spite of the relatively complex procedure of translation cleaning up afterwards.

In a multi-lingual translation memory, all the sentences are coded in TMX, an open XML standard for the exchange of translation memory data created by computer-aided translation and localization tools(Lisa Orgnization, 2005). The data is stored at the sentential level alignment for the translation units, which are marked by a pair of <tuv> and </tuv> and identified by an id number <tu tuid="#">. The following is an example of storing English, German and French TU by the same id="59".

```
<tu tuid="59">
<tuv xml:lang="EN-US">
<seg>There are several methods employed for the production of BSA by commercial manufacturers.</seg>
</tuv>
<tuv xml:lang="de-DE">
<seg>Es gibt mehrere Methoden für die Herstellung von BSA von kommerziellen Herstellern eingesetzt. </seg>
</tuv>
<tuv xml:lang="FR-fr">
<seg> Il existe plusieurs méthodes employées pour la production de BSA par les fabricants commerciaux. </seg>
</tuv>
</tu>
```

Other information about the files and translation units can also be put into a parallel corpus. The file information will be put as the header in the corpus. And the translated sentences will be sent to the corpus and the additional information attached to the sentences will be added to them. The structure of the translation memory with English as the source language and German and French as the target languages will be as follows:

```
<header
creationtool="X"
creationtoolversion="2.8.0"
datatype="unknown"
segtype="sentence"
adminlang="EN-US"
srclang="EN-US"
o-tmf="X"
>
</header>
<tu tuid="0000026"
datatype="Text"
srclang="EN-US"
>
<prop type="x-client">0001</prop>
<prop type="x-client">0001</prop>
<prop type="x-domain">0</prop>
<prop type="x-filename">ABC.doc</prop>
<prop type="x-rowid">0000026</prop>
<tuv  xml:lang="FR-fr"
creationdate="20130518T113038Z"
creationid="Administrator"
 >
<prop type="x-issource">false</prop>
<seg> siège </seg>
</tuv>
<tuv  xml:lang="de-DE"
creationdate="20130518T113038Z"
creationid="Administrator"
>
<prop type="x-issource"> false</prop>
<seg>Hauptsitz</seg>
</tuv>
<tuv  xml:lang="EN-US"
creationdate="20130518T113038Z"
creationid="Administrator"
>
<prop type="x-issource">true</prop>
<seg>headquarter</seg>
</tuv>
</tu>
```

After the structure of the translation memory is designed and the types of information stored in it are decided, a user interface screen can be used to retrieve all the necessary information while the translators are working on the project.

C. Cleaning Up Translation

The target sentence is generated by first looking up the source sentences and calculating their similarities with those in the translation memory. Once the match of the source sentence is found, it will be used in translation in several ways. If no match in the translation memory, manual translation will have to be conducted. Translators have to improve the translation if the match is not to a desirable degree. As the examples in the translation memory build up, chances of finding a match will be easier.

As to the different degree of similarities, the first five closest matches will be put forward for translators' references in a rank. The methods to deal with the referential sentences are as follows:

1) Copy Referential Sentence: If there is 100% similarity between the source and referential sentences, then the referential sentences can be copied for using as the target translation.

2) Manual Translation: If the threshold value of 75% is set as default or changed to another value, sentences to any other similarity degree less than this threshold value will be

considered as unacceptable in translation. It is so that manual translation should be used.

3) Edit the Referential Sentence: Sentences to the similarity degree from the threshold value to the 100% will have to be edited for use, such as adding, deleting or inserting.

In the translation of text in certain specialties or science and technology documents or product catalogues or user manuals or different version of documents etc., fixed sentences and phrasal expression and prefabricated chunks and similar sentences occur in high frequency. Thus, the complete matches or partial matches will account for a large portion of the translation text. The abovementioned methods will provide some useful references for the translation and ensure the quality of the translated text.

IV. Conclusions and Future Work

This research has investigated the key technologies in computer-aided multi-lingual translation memory. The methods to calculate similarities between source sentences and target reference sentences can be approached by several methods, i.e. string-based algorithm, syntax-based algorithm, VSM-based (vector space model) algorithm, word-based algorithm of ontology based on thesauruses or semantic dictionary and that based on large corpus. A minimum edit distance method, based on the Levenshtein distance, may be more applicable to deal with a multi-lingual translation project. Also, an exemplary XML structure of this multi-lingual translation memory is proposed and a demonstration of how the header and body information of translation units at the sentential level alignment is also made within this exemplary XML structure. Then, the methods of cleaning up translation are put forward for a user interface looking-up procedure of finding matches.

The future study should focus on calculation of similarities based on deep structure analysis so as to account for more of the differences among languages and retrieve the closest matches as references. In addition, the multi-lingual term base is not covered and still very important for future study of its uses in translating texts with enormous specialized terms. Last but not the least, translation memory are empty at first and only enlarged along with uses. So the technology to fast expand the translation memory is also worth considering in future work.

References

[1] Dorr, Bonnie J., "Machine Translation Evaluation and Optimization," Part 5, in Joseph Olive, John McCary, and Caitlin Christianson (eds.), Handbook of Natural Language Processing and Machine Translation, 2010.

[2] Lagoudaki, E., Translation Memories Survey, London: Imperial College, from:http://www3.imperial.ac.uk/portal/pls/portallive/docs/1/7294521, 2006.

[3] Makoto Nagao, "A framework of a mechanical translation between Japanese and English by analogy principle," In A. Elithorn and R. Banerji, Artificial and Human Intelligence. Elsevier Science Publishers,1984.

[4] Carl, M. and A. Way., Machine Translation Special Issue: Example-based Machine Translation, vol. 19, 2005.

[5] M. Ramiz, “A new sentence similarity measure and sentence based extractive technique for automatic text summarization,” Expert Systems with Applications, vol. 36(4), 2009:7764-7772

[6] Palakorn Achananuparp, Xiaohua Hu, Xiajiong Shen,"The Evaluation of Sentence Similarity Measures",in DaWaK Proceedings of the 10th international conference on Data Warehousing and Knowledge Discovery, Springer-Verlag Berlin, Heidelberg,2008:305-316.

[7] Sandipan Dandapat, Sara Morrissey, Andy Way, Joseph van Genabith, "Combining EBMT, SMT, TM and IR technologies for quality and scale,"in EACL Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra),Association for Computational Linguistics Stroudsburg, PA, USA,2012:48-58.

[8] Lisa Orgnization,TMX 1.4b specification,Translation Memory eXchange, from:http://www.gala-global.org/oscarStandards/tmx/tmx14b.html, 2005.