International Journal of Applied Linguistics & English Literature
ISSN 2200-3592 (Print), ISSN 2200-3452 (Online)
Vol. 1 No. 5; September 2012 [Special Issue on General Linguistics]

Using Bilingual Parallel Corpora in Translation Memory Systems

Hossein Keshtkar
Linguistics Department
Payame Noor University of Tehran, Iran
E-mail: hkmaram@gmail.com

Tayebeh Mosavi Miangah
Linguistics Department
Payame Noor University of Tehran, Iran
E-mail: mosavit@pnu.ac.ir

Received: 31-07- 2012 Accepted: 15-08- 2012 Published: 03-09- 2012
doi:10.7575/ijalel.v.1n.5p.184 URL: http://dx.doi.org/10.7575/ijalel.v.1n.5p.184

Abstract

Automatic word alignment techniques commonly used in Translation Memory systems tend basically to work at single word level where there is a one to one correspondence between words in subsequences of the two languages. This, results in not being able to fully use subsentential repetitions like clauses, phrases and expressions. In this paper, using spaces between words, a search method named "space-based reduction search" is introduced. The main goal is to maximize the use of parallel corpus resources. We want to show that this search method can significantly enhance the chance of finding matches for subsequences of input sentences; hence applicable in a Sub-Sentential Translation Memory (SSTM) system without running automatic alignment tools.

Keywords: Sub-Sentential Translation Memory, Parallel corpus, Alignment

1. Introduction

As it has been stated by authors and researchers in the field of Machine Translation (MT) and Computer Assisted Translation (CAT), Machine Translation output needs post- edition and is not suitable for publication and releasable purposes. Automatic MT does not allow users to take part in translation process, and after errors occurred, they have to try to fix them. Translation tools can make it possible for users to correct the errors before being transferred into the output. Therefore CAT tools especially TM systems with their interactive environment can help users and translators to a high degree. The most important of these tools are translation memory systems. "… many translators remain convinced that the output of even the best MT systems is not sufficient to facilitate the production of publication-quality texts. To increase their productivity they turn instead to translator support tools. (Macklovitch et al, 2008: 412) The very popular idea in the field of TM systems is to explore a corpus or database of past aligned translations using a software program called TM. Based on this idea, we want to answer the following questions:

1. Is it possible to prepare a parallel bilingual corpus aligned at sentence, phrase and word level with a high level of accuracy?

2. How can such corpus be used as a TM system?

3. Which abilities does this TM system have?

4. To what extent can this TM system automatically perform translation tasks?

5. How well can this TM system do its basic task that is searching and retrieving matches based on common metrics used in the evaluation of CAT tools?

Therefore, we have prepared a parallel bilingual sentence aligned English- Persian corpus using a combination of manual and automatic procedures and want to use it in a TM. As the ability of sentence-level or first generation

TM systems is limited to the rare situations where there are whole sentence repetitions, I hope to find a tool to use the prepared corpus as an SSTM system which is able to retrieve from the database, translations for subsequences of the input sentences. In this paper, using spaces between words as separators, a search method which I call "space-based reduction search" is introduced. The goal is to create interactional environment and maximize the use of corpus resources. I want to know that, to what extent this search method can enhance the chance of finding matches for different subsequences of input sentences. Based on this method the main core or search engine of an SSTM in the form of a software program will be developed in order to test the ability of the search method. This method then will be tested by the software on some parts of the prepared corpus including film subtitles. Recall and Precision scores will be used for evaluation of the method to show that, it significantly enhances the chance of finding matches for subsequences of the searched sentences and is applicable in an SSTM system.

2. Definitions and some historical background

The term "translation memory" has been used to refer to two concepts: 1. a database of past translations in which both source and target text have been aligned at least at paragraph or sentence level and 2. A software program for searching and retrieving parts of past translations for inserting in new translations. (Macklovitch, 2000) addresses these two definitions in more details and gives real examples.

"A TM is essentially a database of previously translated pairs of equivalent source/target language segments (typically, sentences), together with software which, given a source language input to translate, will search in the database and pick out samples which closely match this input. (Whyman & Somers, 1999: 1268)

"Defined most generally, a translation memory is a computerized archive of existing translations, structured in such a way as to promote translation re-use. (Macklovitch, 2000)

"A translation memory system is a type of translation support tool whose purpose is to avoid the re-translation of segments of text for which a translation has previously been produced." (Simard, 2003)

Academic research on the idea of translation memory started in late 70s and the first commercial products were released in late 80s. The original idea is usually attributed to Martin Kay and his 1980 paper although the TM idea is not addressed clearly. "... the translator might start by issuing a command causing the system to display anything in the store that might be relevant to… Before going on, he can examine past and future fragments of text that contain similar material" (Kay, 1997: 19)

(Arthern, 1979) and (Arthern, 1981) explain what we now call a TM system more explicitly: "It must in fact be possible to produce a programme which would enable the word processor to 'remember' whether any part of a new text typed into it had already been translated, and to fetch this part, together with the translation which had already been translated … Any new text would be typed into a word processing station, and as it was being typed, the system would check this text against the earlier texts stored in its memory … One advantage over machine translation proper would be that all the passages so retrieved would be grammatically correct. In effect, we should be operating an electronic 'cut and stick' process which would, according to my calculations, save at least 15 per cent of the time which translators now employ in effectively producing translations." (Arthern, 1981: 318)

This idea was used in Alps one of the first commercial systems developed in Birmingham Young University which was called 'Repetitions Processing' and was able to retrieve only exact matches. As (Melby, 1995) says, the source code of this system was then used by IBM in its well known system, Translation Manager.

3. Review of the methods directly related to SSTM systems

One criterion for classification of TM systems is whether the system can perform at sentence level or sub sentence level so TM systems are often classified as sentence level or first generation systems and sub sentence level or second generation systems. The ability of the first category is limited to retrieving the whole sentences in the rare situations in which sentences are repeated, for example translation of revised documents or texts which are repetitive in nature. Because of these limitations some methods have been proposed by researchers trying to develop second generation system that is SSTM system. Although these methods have been proposed for the alignment and retrieval of multiple word subsequences which are the fundamental task of these systems, but many of these methods tend to work only at single word level where there is a one to one correspondence between the elements of the tow languages involved.

Evaluations of these methods show that although they highly increase the Recall (R) but they highly decrease Precision (P) too. Single word alignment and retrieval as will be shown in 4.2 of this paper will basically lead to low P. On the other hand clauses, phrases and especially multiple word expressions and idiomatic constructions which are called fixed expressions too, compose a large proportion of written and spoken languages as Mel'čuk says: "Fixed expressions are crucial because a large part of what we say and write is made up of such expressions, rather than separate words" (Mel'čuk 2001: 24).

What they have in common is using probabilistic approaches for the alignment and retrieval of the subsentential strings which may be single word or multiple word subsequences of a sentence. In other words these methods use a statistical translation model under different conditions or constraints to find the most probable alignment that is finding candidate segments within the target text for the translation of a segment in the source text.

Translation spotting which is as (Simard, 2003) says the byproduct of word level alignments is a term coined by (V´eronis and Langlais, 2000). It is used for the task of identifying the words in a target language translation that correspond to some given words in a source language text. Translation spotting which has been used in SMT and in SSTM is based on some probabilistic models. There are different models of translation spotting which are mentioned here very shortly as our purpose is not to explain them.

Viterbi translation spotting links one source word to one target word. The Viterbi algorithm which was first described in (Viterbi 1967) is based on the maximum likelihood alignment.

Post-processings are used because the answers produced by Viterbi translation spotting are not necessarily contiguous. This model tries to make the answers contiguous that is to edit them to some extent using processings like *expansion* and *longest-sequence* so that they correspond to natural word order as much as possible or *zero-tolerance* to discard problematic answers. Other examples are *Contiguous* translation spotting and *Compositional* translation spotting.

Based on these models some variations have been proposed and some of them are widely used in SSTM and SMT for translation spotting. For example in IBM-style alignments, IBM models 1 to 5, a single target word can be connected to several source words. Alignment models proposed by Melamed (1998) and Wu (1997) allow "one-to-one" alignments. Planas (2000) proposes an approach for an SSTM which is based on sequences of syntactic chunks, as defined by Abney (1991). The contents of the TM and the new text are segmented into chunks; sequences of chunks from the new text are searched in the TM and the translation of the matched sequences will be proposed to the user as partial translations of the current input.

4. Suggested method

As it was said in the previous part translation spotting models proposed are based on finding the most probable alignment and tend to work in single word level, so produce high R but low P. In addition, using constraints like contiguity and compositionality have led to better results mainly in R. But according to (Simard, 2003) "…precision is possibly more important than recall in a TM application."

In SSTM systems analogous to what human translators do, that is translate a text sentence by sentence, the input source fragments are mainly sentences which are feed one by one to the systems automatically or by the user. The systems then usually divide them into sub sentences or segments. These segments do not usually correspond to the boundaries of grammatical phrases which compose the sentence. So if we suppose that thanks to some remedies made to the probabilistic alignment methods as mentioned above, the systems can recall usable matches for the segments which corresponded to grammatical phrases but what is recalled for the other segments needs much more edition. In other word although R will be high but edit distance will be high because of low P and it will take much time and effort for the user to fit recalled matches for final translation. So it seems that the overall function of an SSTM depends heavily on probabilistic alignment methods and the method of segmentation of the input sentences.

As we know TM systems like MT systems perform better when they are used for domain specific texts, so if we can find a method to train the used algorithms to calculate all possible subsequences of every input sentence and search each of them in a domain specific storage of bilingual parallel aligned corpus, it is more likely to find usable matches with high P and low edit distance, without using probabilistic alignment methods. For this purpose we need a relatively large storage of domain specific parallel bilingual sentence aligned texts in which sentences have various lengths. Such storage should be enriched with subsentential strings or multiple word expressions, for example clauses, phrases, idiomatic expressions and two word constructions.

There are at least two reasons why it is more likely to find usable matches with high P and low edit distance in such a situation:

1. Single word subsequences produced and searched by the algorithm will usually return no answers since we do not incorporate single words in the records in our database, so the absence of single word matches will lead to higher P.

2. Multiple word subsequences which are not grammatical or does not correspond to phrase boundaries but are produced and searched by the algorithm will return no answer as we have not incorporated such below sentence level strings.

In other words what is recalled by the system will more likely to be 1. Subsequences having two word or more, 2. subsequences which usually are grammatical and more natural so that the user need do less effort in the form of editing, deleting, inserting, reordering to fit the translation.

The chances are that the result will be high P and less edit distance; what SSTM system providers are trying for. If successful, this method would not necessarily need statistical calculation for probabilistic alignment methods; hence would not need running automatic alignment tools most notably Giza++. Instead we need to enrich our corpus as much as possible. Of course products of automatic alignment tools in the form of aligned subsentential strings≧ two words, controlled by qualified translators can be or must be incorporated in the corpus as an enriching ingredient.

To find such an algorithm it seems to be useful to look more closely at visible aspects of segmentation and alignment in SSTM systems. Some criteria which are used in GUI of these systems for users in the form of system settings to enable them to use underlying probabilistic methods for alignment and segmentation are more visible. Examples of these visible criteria are punctuation marks, embedded formatting clues which for example mark the start and end of a paragraph, tag sets and phrase structure rules which are used in POS taggers and morphological analyzers.

Using separators like punctuation marks or formatting clues used and embedded in electronic texts are relatively useful for sentence level alignment. Differences in using punctuation marks among languages, structural differences and unavoidable changes in the number and order of sentences in target language texts are among the reasons why these separators are not fully reliable. On the other hand using phrase structure rules, POS taggers, morphological analyzers for segmentation, alignment and reordering retrieved matches are highly language dependent. Automatic alignment techniques have remarkable functions mainly in two situations: First, at paragraph and sentence level and second, at single word level where there is a one to one correspondence between words in subsequences of the two languages. This, results in not being able to fully use subsentential repetitions like clauses, phrases, fixed expressions, multi word expressions and idiomatic expressions.

Pressing space key and inserting one space between words in sentences in most of languages is consistent and follows a clear pattern. Based on this, I propose a method for searching source sentences in a TM database stored in advance with a relatively large amount of sentence level aligned parallel texts. Then based on this method, a framework for development of a software program for searching such corpus is introduced. This software program which I call "HKTM" Ver. 1.2 is then used as the main core or search engine of an SSTM.

4.1 Detailed introduction of the method

Based on the strategy used in this method I call it "Space Based Reduction Search". This method and its related software program have been built on four maxims:

1. Maximum or optimized use of corpus resources.

2. User-machine interaction.

3. Preventing error transition into the output.

4. Controlled and accurate subsentential alignment during translation.

As it will be very expensive and time consuming to develop a full SSTM, in order to evaluate the proposed search method I have prepared a rough version of the software program which will be used as the main core or search engine of the SSTM. In this evaluation, given that there is no sentence level match for the searched sentence, we want to know that to what extent this search method can enhance the chance of finding matches for subsequences of that sentence.

Most of the automatic word alignment techniques for example Giza++ use co-occurrence statistics and usually align only one source word to one or more target words and so are not useful for aligning clauses, phrases and idiomatic expressions where there is not a one to one correspondence between the source and target elements.

Although through some changes and combination with other methods, for example bi-directional running of Giza++ better results have been achieved, but they have been criticized both for difficulties in compiling and running and their output especially for purposes like TM systems. So these word alignment tools although useful for many purposes, does not ensure the use of clauses, phrases and idiomatic expressions even when they exist separately (not within a sentence) in the database.

Using this search method we are sure that all possible subsequence of a sentence will be searched and if there is a match it will surely be retrieved and showed to the user for edition or verification in interactive mode or inserted directly into the target text in automatic mode. So this search method maximizes the use of corpus resources.

To describe "space based reduction search" we suppose that a sentence is a set. This search method uses a simple algorithm which makes it possible for the system to show all possible sub-sets of the set which are all possible subsequences of the sentence. The searched sub-sets have these two characteristics: 1. Sub-sets which have only one element. 2. Sub-sets which have at least two sequential elements (according to word order of the sentences).

If we show a five word sentence in the form of a five element set as bellow:

{Computers can highly assist translators.}

{ 1 2 3 4 5 }

The sub-sets which will be searched are 15 sub-sets which cover all possible subsequences of the sentence and will be searched by the algorithm as follows:

1. {1 2 3 4 5} {Computers can highly assist translators}

2. {1 2 3 4} {Computers can highly assist}

3. {1 2 3} {Computers can highly}

4. {1 2} {Computers can}

5. {1} {Computers}

6. {2 3 4 5} {can highly assist translators}

7. {2 3 4} {can highly assist}

8. {2 3} {can highly}

9. {2} {can}

10. {3 4 5} {highly assist translators}

11. {3 4} {highly assist}

12. {3} {highly}

13. {4 5} {assist translators}

14. {4} {assist}

15. {5} {translators}

Based on this search method all possible subsequences of a sentence are searched in the database and it is possible for the user to edit, reorder, delete or accept the retrieved segments and the program links or aligns finally accepted segments and stores them in the database as matching pairs.

To retrieve matches below sentence level and above single word level, this search method does not depend on automatic word alignment techniques; instead it relies on the richness of the corpus.

Advantages:

1. It is based on spaces between words as separator so it is consistent, simple and does not need complicated algorithms and statistical calculations.

2. It is language independent and can be used for other languages bi-directionally.

3. It maximizes the use of corpus resources which are usually prepared with great difficulties.

4. It provides an interactional environment to ensure precision and accuracy and prevent error transferring into the output.

5. It does not necessarily depend on grammatical rules and structures so there is no need to formalize the rules.

6. It does not necessarily need running automatic alignment tools.

Disadvantages:

1. Because it does not use structural rules, in automatic mode, the word order in target sentences will be much like the source language word order.

2. When it finds a match for a subsequence, inserts it into the target sentence and leaves other subsequences which may be better for translation.

The first problem that is the word order can either be corrected by user-program interaction which is a principle in this method or can be partially solved by incorporating sufficient phrase structure rules or word order patterns of desired target language.

The second problem can be targeted by tuning the program so as to show retrieved matches for all possible subsequences, rank them if needed and allow the user to choose the best one for inserting into the target text.

4.2 Testing and Evaluation

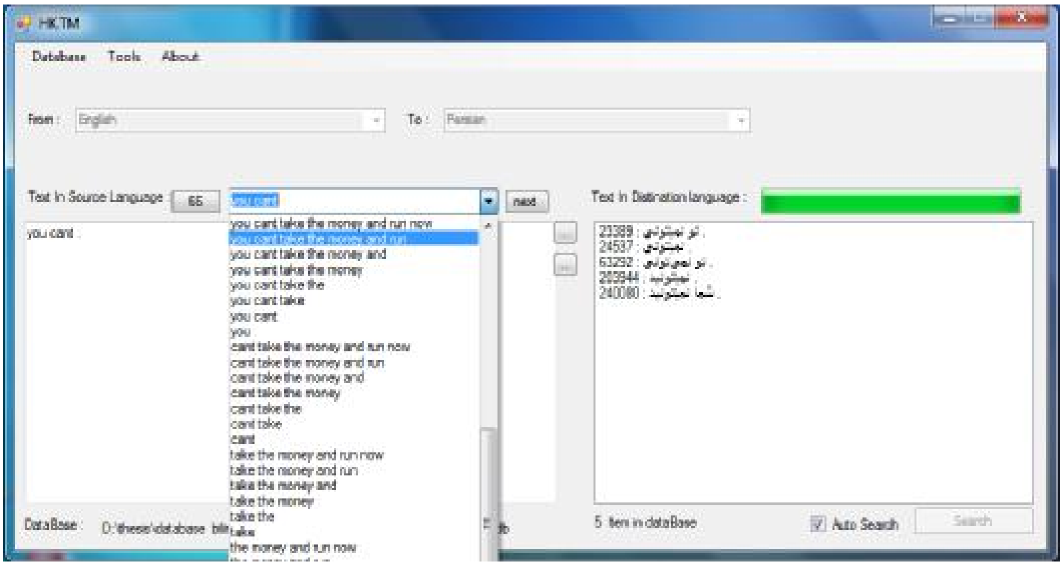
The software developed to use the proposed search method as the base of an SSTM system is now able to connect to the database, show all the possible subsequences, searching based on the method and showing the results both automatically and interactively. Given that there is no exact match for a sentence, to ensure that, this search method can significantly enhance the chance of finding matches for possible subsequences of that sentence, a test set was prepared. Since, a large proportion of the texts in the prepared corpus are film subtitles, and given that a TM system better perform in specific domains, the test set was chosen from film subtitles. One hundred sentences were chosen from film subtitles and were carefully sentence-aligned and saved as reference set. Then, sentence by sentence, the one hundred source sentences searched in the database using "Space based Reduction Search" by “HKTM” VER .1.2.

Sentence level matches were excluded and the results were recorded in two situations. First situation: found matches for subsequences at single word level up to below sentence level. Second situation: found matches for subsequences above single word level (two words or more) up to below sentence level.

In the first situation total number of retrieved matches for total subsequences of the one hundred sentences searched in the database was 454 among which 290 matches were usable without edition (reordering was needed). So, %63.87 percent of total retrieved matches were usable in translation.

In the second situation, total retrieved matches for the possible subsequences containing two words or more in the one hundred searched sentences in the database was 108 among which 83 matches were usable without edition(reordering was needed). So, %76.85 percent of them were usable in translation.

The following pictures are two screen shots of HK.TM version 1.2. The first one shows all subsequences of a searched sentence and the retrieved answers for the chosen subsequence. The second screen shot shows other details about the software.



From : English

To : Persian

Text In Source Language : 65 you cant

you cant .

you cant take the money and run now
you cant take the money and run
you cant take the money and
you cant take the money
you cant take the
you cant take
you cant
you
cant take the money and run now
cant take the money and run
cant take the money and
cant take the money
cant take the
cant take
cant
take the money and run now
take the money and run
take the money and
take the money
take the
take
the money and run now

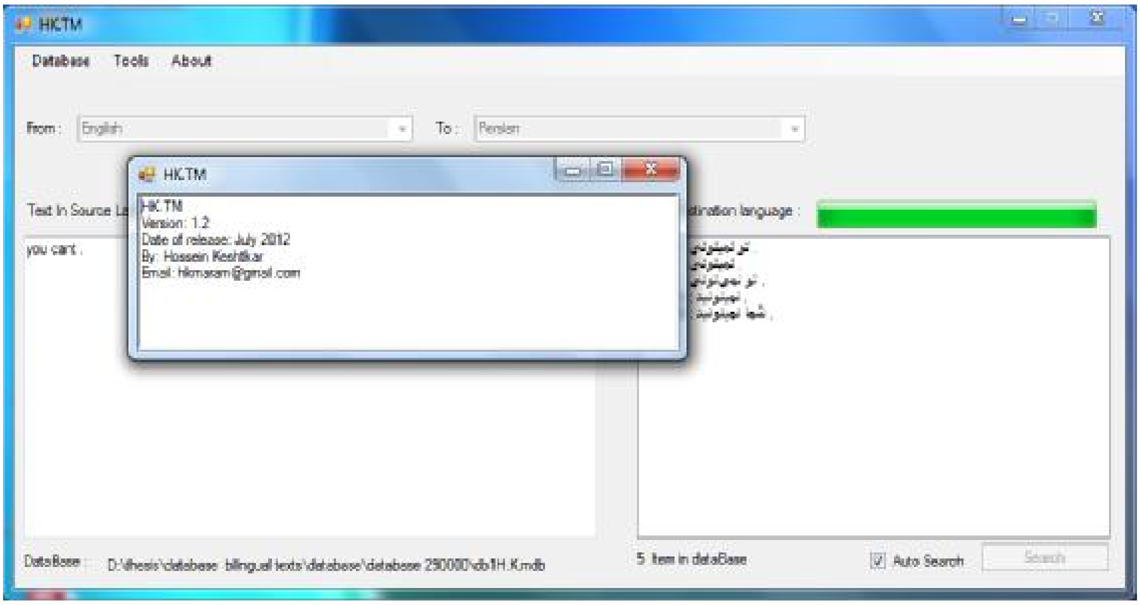
Text In Distination language :

23389 : تو نمیتونی .
24537 : نمیتونی .
63292 : تو نمی‌تونی .
203944 : نمیتونه .
240080 : شما نمیتونید .

5 item in dataBase

☑ Auto Search

DataBase : D:\these\database bili



HKTM

HK TM
Version: 1.2
Date of release: July 2012
By: Hossein Keshtkar
Email: Hkmaram@gmail.com

you cant .

تو نمیتوانی .
نمیتوانی .
تو نمی‌توانی .
نمیتوانید .
شما نمیتوانید .

DataBase : D:\thesis\database bilingual texts\database\database 290000\db1H.K.mdb

5 item in dataBase

☑ Auto Search

These results show that: 1. The chance of finding matches for subsentential strings in a domain specific corpus using "Space based Reduction Search" is high, without running word alignment tools. 2. The number of retrieved matches is remarkable. 3. In the second situation that is when one word matches are excluded, higher percentage of retrieved matches are usable in translation; so that a 13 percent increase in usability is observed compared to the first situation where one word subsequences are also taken into account.

Here, two common metrics are used for close evaluation of the method. As TM systems are to some extent similar to Information Retrieval (IR) systems, common metrics used for evaluation of IR systems are widely used in TM systems evaluation. Among these metrics, "Recall" and "Precision" hereafter referred to as R and P are commonly used. For example (Whyman & Somers, 1999: 1268) site the similarities and differences of IR and TM systems and use these metrics in their evaluation of TM systems. These metrics are interrelated so they affect each other. "It is noted that P and R are not independent: an improvement in one is generally at the expense of the other, that is higher precision will generally lead to less item being retrieved, hence lower recall, and vice versa." (Whyman & Somers, 1999: 1271) Every one of these metrics is calculated in the form of a fraction. "Recall is calculated as the fraction of relevant documents found among all relevant documents whereas precision is the fraction of relevant documents in the result set." (Mandl, 2008: 28)

Based on these facts R and P are used for close evaluation of the proposed method. Here, R is defined as the fraction of found items among expected items (an item is each subsequence of the searched sentence with its retrieved match or matches) and P as the fraction of usable items among found items.

In the first situation, to lessen the effect of single word matches in the calculation of R and P, It is expected that for each search sentence, at least two items will be found. So it is expected that 100 items be recalled or retrieved. Also, very frequent small words such as "and", "for", "is", "to" will be excluded.

After searching, in 87 out of 100 searched sentences, matches having above characteristics were found. So, R is 0.87.

R_1 = found items ÷ expected items = 87 ÷ 100 = 0.87

And, in 66 out of 87 sentences, the found items were usable in translation of the searched sentences without edition (obviously reordering was needed). So, P is 0.75.

P_1 = usable items ÷ found items = 66 ÷ 87 = 0.75

In the second situation, it is expected that for each sentence at least one item for two word or above two word subsequences will be found. So it is expected that 100 matches be retrieved. After searching, in 55 out of 100 searched sentences, matches having above characteristics were found. So, R is 0.55.

R_2 = found items ÷ expected items = 55 ÷ 100 = 0.55

And in 49 out of the 55 sentences, found items were usable in translation of the sentences without edition (obviously reordering was needed). So P is 0.89.

P_2 = usable items ÷ found items = 49 ÷ 55 = 0.89

These scores also show that:

1. Searching based on "Space based Reduction Search" method in a bilingual parallel sentence aligned domain specific corpus makes it highly possible to find matches for the subsequences of searched sentences.

2. Obtained scores both for R and P in both situations are above 0.50. So this search method significantly enhances the chance of finding subsentential matches in such corpus. It is worth reminding that no automatic word alignment tool has been run.

3. In the second situation where two word and above two word subsequences were recorded, R is lower than in the first situation but the usability of recalled items in translation is very high; hence precision is higher than recall.

4.3 Results

Subsequences discussed above will be referred to as translation units hereafter. When translation units composed of two words or more, are retrieved and used in translation, edit distance will be low. It can be concluded that although automatic word alignment tools for example Giza++ are helpful but it is not necessarily needed to use them in the proposed method and the SSTM system which will be built based on it. There are at least two reasons. First, compiling and running such programs, especially on large corpora like the corpus used in this research, is very time consuming and expensive. Second, these techniques mainly work at single word level and cannot considerably help in the situation where there is not one to one correspondence between the elements of the source and target strings; for example in clauses, phrases and idiomatic expressions. So, although word alignment tools increase R but at the same time they decrease or at least do not improve P. When translation units are chosen in the form of two word segments or more, items which are retrieved are not too short and P is higher than R (here, 0.89 vs. 0.55 respectively).

Therefore, to achieve a high level of P, translation units are to be retrieved should not be very small; that is they should be at least two words or more. These findings are also in total accordance to the results found by (Simard and Langlais, 2001) regarding translation units and their effects on P and R.

5. Conclusion and further development

Now, Machine Translation output needs post edition and is not suitable for publication and releasable purposes. Therefore CAT tools especially TM systems with their interactive environments can highly help users and translators. To summarize and conclude, I return to the paused questions in introduction part and answer them.

5.1 With regard to computer development, data storage instruments, data mining information retrieval techniques and huge resources of bilingual texts accessible via internet it is possible to prepare large scale parallel corpora. There are automatic and semi-automatic methods for mining, purification, edition and preparation of the texts and storing them in suitable database formats. There is also software program for alignment at paragraph and sentence level. The problem mainly arises in subsentential levels. Most of the software programs introduced for subsentential level alignment for example Giza++ work at single word level and in the situation where there is one to one correspondence between the elements in source and target language; so are not suitable for the alignment of clauses, phrases and idiomatic expressions. As a result of this fact and based on experiences from prepared corpus, to ensure precision and accuracy of the aligned corpus a combination of manual and automatic or semi automatic methods will be helpful for the preparation of parallel bilingual sentence aligned corpora.

5.2 TM is usually used to refer to two concepts, one as a database of bi-text of previously translated text and the other as a software program usually called a TM system with a user interface enabling translators to use inter and intra–text repetitions for the texts to be translated. TM systems, because of the nature of their segmentation and alignment techniques, mainly give relative results in two situations: first, when there is a sentence-level match, second, when a user or translator entirely does his or her translation within the interactive environment of a TM system from the beginning, and stores them. In this situation after a learning curve, the user can use the increasing inter and intra –text repetitions.

So, most useful subsentential repetitions for example clauses, phrases and idiomatic expressions available in past translations are not used in TM systems. Based on this fact, the proposed search method which in fact is the basis or search engine of a TM system was introduced. Among its basic principles the two most important principles are: first, providing high level of interaction between the translator and system to ensure precision and accuracy and preventing errors from being transferred into the output and second, maximizing the use of corpus resources prepared spending time and money. Needed functions and facilities for using the software as an SSTM system has been discussed, predicted and can easily be added to the search engine.

5.3 The software which is the core or search engine of an SSTM system is able to automatically perform the "Space based Reduction Search". This method makes it possible to search all the subsequences of the input sentences based on the second principle mentioned in 5.2

5.4 Showing all the subsequences of input sentence, searching the database in a very short time and showing the matches are done automatically by the software. It is possible for the user to take part in each phase based on the first principle mentioned in 5.2.

5.5 Results obtained and mentioned in evaluation part showed that the ability of the method for searching and finding matches is remarkable. Close examination based on recall and precision scores showed the effectiveness of this method for finding usable matches for subsequences below sentence level and above single word level in a domain specific corpus.

It is important that these scores obtained without running any automatic alignment tool. It is also worth considering that precision is possibly more important than recall in a TM application. So, based on these findings and discussions it can be concluded that combining the proposed search method and the prepared corpus will produce basic requirements for the development of an SSTM. For more convenience the method can be performed on the other language domains available in the corpus. After adding other functions of an SSTM system, the overall performance of the system can be evaluated using other metrics such as edit distance. These are among future works to be done.

References

-Abney, S. (1991). *Parsing by Chunks.* In: R. C. Berwick (ed.), Principle-Based Parsing: Computation and Psycholinguistics, 257–78, Dordrecht: Kluwer.

-Arthern, P. J. (1979). *Machine translation and computerized terminology systems: a translator's viewpoint.* In: B.M. Snell (editor), Translating and the Computer: Proceedings of a Seminar, North-Holland, 1979, pp. 77–108.

-Arthern, P. J. (1981). *Aids unlimited: the scope for machine aids in a large organization.* Aslib Proceedings, 33, 309–319 (1981).

-Kay, M. (1997). *The proper place of man and machine in language translation.* In: machine translation, volume 12, Nos. 1-2, 1997, 3-23 (reprint from 1980)

-Macklovitch, E. (2000) *Two Types of Translation Memory.* In: Translating and the Computer 22: Proceedings from the Aslib conference held on 16 & 17 November 2000,

-Macklovitch, E., Lapalme G. & Gotti F. (2008). *TransSearch: what are translators looking for?* AMTA-2008. MT at work: Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas, Waikiki, Hawai'i, 21-25 October 2008; pp.412-419.

-Mandl, T. (2008). *Recent Developments in the Evaluation of Information Retrieval Systems: Moving Towards Diversity and Practical Relevance.* Informatica 32 (2008) 27–38.

-Manning, C. D. & Schiitze H. (1999). *Foundations of Statistical Natural Language Processing.* The MIT Press Cambridge, Massachusetts, London, England.

-Melamed, I. Dan. (1998). *Word-to-Word Models of Translational Equivalence.* Technical Report 98-08, Dept. of Computer and Information Science, University of Pennsylvania, Philadelphia, USA.

-Melby, A. K. (1995). *The Possibility of Language: A Discussion of the Nature of Language.* John Benjamins, 1995, p. 225f.

-Simard, M. (2003). *Translation Spotting for Translation Memories.* HLT-NAACL 2003 Workshop, *"Building and using parallel texts: data driven machine translation and beyond"*, 31 May 2003, Edmonton, Canada.

-Simard, M. & Langlais, P. (2001). *Sub-sentential Exploitation of Translation Memories.* MT Summit VIII: *Machine Translation in the Information Age*, Proceedings, Santiago de Compostela, Spain, 18-22 September 2001; pp.335-339.

-V´eronis, J. & Langlais, P. (2000). *Evaluation of Parallel Text Alignment Systems – The ARCADE Project.* In Jean V´eronis, editor, *Parallel Text Processing*, Text, Speech and Language Technology. Kluwer Academic Publishers, Dordrecht, The Netherlands.

-Viterbi, A. J. (1967). *Error bounds for convolutional codes and an asymptotically optimum decoding algorithm*. IEEE Transactions on Information Theory IT-13: 1260-269.

-Whyman, E. K. & Somers, H. L. (1999). *Evaluation Metrics for a Translation Memory System.* SOFTWARE-PRACTICE AND EXPERIENCE *Softw. Pract. Exper.,* 29(14), 1265–1284 (1999)

-Wu, D. (1997). *Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora*. Computational Linguistics, 23(3):377–404, September.