Fuzzy Match Score and Translation Memory Match: A Linguistic Insight

Sandipan Dandapat[1], **Shakuntala Mahanta**[2], **Sibansu Mukhopadhyay**[3]
[1] Department of Computer Science and Engineering
[2] Department of Humanities and Social Sciences
[1,2] Indian Institute of Technology Guwahati, Assam India
[3] Society for Natural Language Technology Research, Kolkata, India
{sdandapat,smahanta}@iitg.ernet.in, sibansu@gmail.com

Abstract

Fuzzy match score (FMS) is the most widely used metric for finding similar examples from a translation memory database for an input string. Source-language FMS is considered to be the indication of the amount of target edits required by a human post editor in the translation process in a computer aided translation (CAT) system. In this paper, we conduct a detailed study of the inadequacies of the standard FMS. Furthermore, we look into the linguistic parameters for identifying the inappropriateness of FMS-based similarity measure. Finally, we propose some possible ways of incorporating linguistic features into the FMS computation to extract better candidates in a CAT-based translation framework.

1 Introduction

Translation memory (TM) technology is widely used to assist the translation process within industrial language service providers' (LSPs) localization work-flow. LSPs often use computer aided translation (CAT) systems to assist professional translators. A CAT-based system (Bowker, 2002) segments the input text to be translated, and compare each input segment with the TM database to find one or more *close* target equivalents for the input source segment. Professional translators select one of these close target equivalents and produce the desired translation with modifications.

The back bone of a TM-based technology lies in finding the closest possible match from a TM-database D, for a given input segment (s'). The TM-database D, consists of translation pairs $\langle s, t \rangle$, where s is a source language segment (typically a sentence) and t is its translation in target language. TM-based technology uses a *similarity function* f to find a set of translation units $\{(s_i, t_i)\}$ from D. Fuzzy match score (FMS) (Sikes, 2007) is the most commonly used measure for finding such a similarity.

In order to translate a new document, LSPs use a CAT-based system to find the closest match $\langle s_i, t_i \rangle$ for each input segment s', and edit t_i to produce the desired translation t'. The principle of using a CAT-based system is to reduce the number of edits by selecting a segment (close match s_i) and modifying its target equivalent t_i instead of translating s' from scratch. This is quite effective for LSPs in terms of time and money. However, in a CAT-based system, the similarity function f measures the similarity between two source-language strings while the objective is to reduce the number of edits in the target language text. This is based on the assumption that if two sentences are quite close in one language their translations will also be closer in target language. This assumption can be misleading and may choose some $\langle s_i, t_i \rangle$ pair where source language similarity is highest but may not be efficient in relating the required lowest possible target language edits.

In this paper, we address the issue of FMS for modelling the similarity function in a CAT-based system. We study different linguistic aspects that are not captured in traditional FMS-based similarity. We propose from our linguistic study, how FMS similarity can be improved using different linguistic phenomenon. We also propose a method for incorporating language specific information into the state-of-the-art FMS-based similarity function to estimate better target language similarity.

The rest of the paper is organized as follows. Next section presents some background for our work. Section 3 describes the detailed working principle of TM-based technology using FMS. In section 4, we describe issues with FMS and report our linguistic observation to handle those issues. In section 5, we propose some alternative computational treatment of FMS when used in a CAT-based system. We conclude in section 6 with some avenues for future work.

2 Related Work

The different approaches of machine translation (MT) can be primarily classified as either rule-based or data-driven. Although they represent different approaches to MT, today they borrow ideas heavily from each other. Today, the field of research in MT is largely dominated by data-driven, or corpus based approach. Example-based MT (EBMT) and statistical MT (SMT) represents the two threads of what is known as data-driven MT, with SMT, by far, being the most prevalent of the two.

The above paradigms of MT represent a fully automated end-to-end translation procedure. However, the use of CAT-based tool, a semi-automated approach of translation is becoming popular due to its on going success in assisting LSPs work flow. The heart of the CAT-based system is the TM-database. A TM essentially stores source- and target-language translation pairs (called *translation units*, TU) for effective reuse of the previous translation. The concept of TM (Kay, 1980)is often linked with the concept of EBMT. EBMT is a fully automated approach that uses TM-database to find closely matching sentences for the source-language sentence to be translated. After retrieving a set of example, with its associated translations, EBMT systems automatically extracts translation of suitable fragments and recombine them to produce the desired translation. On the other hand, a CAT environment is a semi automated approach. First, the CAT-based system automatically retrieves the closest match for each input segment to be translated. Furthermore, professional translators select and recombine (with modification) the translation of each input segment based on the TM match.

Recent research related to the CAT system using TM technology primarily focus on three aspects:

- Fast and efficient source-segment searching in the TM-database
- Guidance for target side change
- Combining MT and TM

The first factor affects the runtime performance of a TM. Traditionally, the TM search uses quadratic time complex FMS-based similarity. This is quite time consuming when the size of the TM database is large. TM users need to find the best match from the TM-database in real time. This area is still under active research with a few recent efforts, e.g. Koehn and Senellart (2010a) used an *n*-gram-based matching method to find the potential candidate from a large database. Then A*-search was applied to filter some candidates and finally used A*-parsing to validate the matched segment. Their method outperform the baseline by a factor of 100 in terms of the speed and look up time.

Dandapat et al. (2012) used an IR-based indexing technique to speed up the quadratic time-consuming matching procedure. They showed that index-based matching procedure substantially improves the search time without affecting the translation quality. Furthermore, Laveling et al. (2012) conducted a detailed comparison of different measures that can be used for approximate string matching in an EBMT/TM framework.

The second aspect of research focuses on giving some hints for changing the target segment t_i to obtain the desired translation t'. This area mainly tries automatic modification (selection and recombination of segments) of t_i. Koehn and Senellart (2010b) used TM to retrieve matches for input segments, and replaced the mismatched parts using an SMT system to fill the gaps in the target-side. Zhechev and Genabith (2010) used a sub-tree alignment technique to align source–target pairs from the TM to detect gaps with the new input segment and used the SMT system to fill those gaps to maximize performance. Some recent work has explored the possibility of marking possible changes in the target segment t_i to assist the human translator. Espla et al. (2011a) used word alignment to predict which target words have to be changed and which should be kept unedited. They showed that their approach worked with high precision for higher FMS. Furthermore, Espla et al. (2011b) computed the alignment strength using an MT system to provide the target-language edit hints.

In the third direction, recent research tries to integrate TM and MT system together to use the best of the individual system. He et al. (2012) have successfully shown that integration of TM and MT outperforms the individual system and more effective in terms of time and money (Dara et al., 2013). More recently, Wang et al. (2013) have proposed integrated models to incorporate TM information into phrase-based SMT and the proposed models achieved significant improvement in translation accuracy.

All the above research rely on FMS to find the closest match pair $\langle s_i, t_i \rangle$ for a given input segment s'. The research primarily focus on aligning t_i with s' to provide guidance to human translators. However, to the best of our knowledge, no one has tried to improve the similarity measure to find better $\langle s_i, t_i \rangle$ candidate pair. FMS-based similarity only uses source-language similarity based on surface words. No work has used target language information during similarity match although the objective is to reduce the target language edits. We attempt to study the linguistic aspects of a source–target language pairs and try to incorporate those linguistic information into the state-of-the art FMS score to find better candidate matches.

3 Translation Memory and FMS

A TM is essentially a database that stores source- and target-language translation pairs for effective reuse of previous translation. When a new sentence is to be translated, a TM engine retrieves an entry from the database whose source side is the most similar to the input strings and present to the human translators. The similarity between in input string (s') and the source side TUs in the TM is often calculated using the edit-distance-based (Levenshtein, 1965) FMS as in (1):

$$\text{FuzzyMtach}(t_i) = 1 - \min_{s_i} \frac{\text{EditDistance}(s', s_i)}{\max(|s'|, |s_i|)} \quad (1)$$

where, s' is the source-side segment to be matched with the TM, s_i is a TU in the TM and t_i is the TM hit based on fuzzy-match score.

If a TU in the TM matches the input segment exactly, the translation of this TU can be directly reused without any further processing. In the case of partial matching, the translation is extracted from the database as a skeleton translation which is post edited by a human translator to produce the correct translation.

During the process, it is often the case that a larger number of TU gets similar FMS and there are ties in the highest FMS. We conduct a study to estimate the ties in the highest FMS using IWSLT[1] English–Turkish corpus.[2] We construct the TM database using 22k English–Turkish translation pairs and compute the FMS for 386 English sentences from IWSLT'09 test set. We find that around 58% sentences are ambiguous with respect to highest FMS-based retrieval from the TM. Figure 1 shows the detailed statistics of ambiguity in FMS. We can see that a large percentage of sentences have 2, 3 or 4 sentences from the TM with same highest FMS.

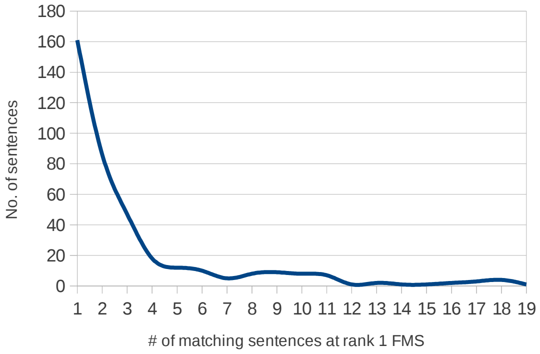

Figure 1: Ties in highest FMS: X axis indicates the level of ambiguity and Y axis denote the number of sentences with a particular ambiguity level.

In the case of a tie, the TM engine retrieves a random TU pair which belongs to the set of TUs with highest FMS. This random selection often leads to a selection that require more edits compared to another competing TU pair with same FMS. Consider the example in Table 1.[3]. Although both the sentences have equal FMS score but they required different amounts of edits to obtain the desired translation (t') for the given input s'. The portion marked in red indicates the required edits in t_i. This clearly shows that random selection may lead to a less edit-effective choice. However, we can use linguistic information in order to handle some such cases and produce a better edit-effective t_i to assist human translators. In the following section, we provide examples of cases from Assamese and Bangla where tra-

[1] International Workshop on Spoken Language Translation. http://mastarpj.nict.go.jp/IWSLT2009/2009/12/evaluation-campaign.html

[2] Note that the target language is independent for this observation as we measure the FMS in source (English) side.

[3] We present all Indian Language examples in ITRANS (Chopde, 2001) notation

Table 1: Example with a tie in the source-side FMS.

Input(s'): I ate spaghetti in the morning .		
s_i	FMS(s_i,s')	t_i
They ate spaghetti in the morning .	0.86	tArA sakAle spaghetti kheYechhila .
I ate *apple* in the morning .	0.86	Ami sakAle Apela kheYechhilAma .

ditional FMS-based similarity fails to identify the most effective TU separately.

4 Linguistic Issues with FMS

In order to instantiate how FMS may falter when source sentences with higher similarity does not lead to proportional similarity in the translated sentences, we choose some sentences with the highest source language similarity and show that their target language edits have a wide range of dissimilarities. The linguistic examples which have been taken up for this purpose belong to a widely divergent set. While some examples are simple declarative sentences, others are complex with one or more subordinating clauses. We study English-to-Assamese and English-to-Bangla translation directions in our work.

4.1 Case and Agreement

We first take a few instances of case and agreement which demonstrate source and target differences. Case and agreement are fundamental aspects of human languages. Among these, case encodes a grammatical relation between two constituents in a clause. A grammatical relation is a type of functional or semantic relation encoded in one constituent in relation to another constituent. Such grammatical relations are usually indicated by certain morphemes. Therefore morphological elements which are borne by nominal elements would normally indicate the grammatical relation they bear with the predicate. Grammatical cases could be nominative, accusative, dative and genitive. Oblique cases, on the other hand mark a semantic relation rather than a grammatical relation. In English, oblique case is always marked by a preposition.

Agreement can be understood as a process where a grammatical element changes in relation with the features of another grammatical element. There are various types of agreement phenomena reported in the literature, namely subject-verb agreement, object-verb agreement, adjective-noun agreement etc.

4.1.1 Subject-Verb Agreement

In this type of grammatical agreement, the verb changes its Φ features (also called PNG - person, number and gender features) based on the properties of the subject NP. In the languages concerned here, English verbs show minimal agreement with the Φ features mentioned above - the only agreement which is surface apparent is the third person singular. In contrast, in the target languages taken up for discussion here, the verb always agrees with the person feature of the subject NP. The example below shows this effect in translation.[4]

1. (a) s': *I eat (rice).* ⇒ *ma;i bhAta khA.N* .
   (b) $\langle s_1, t_1 \rangle$: **We** eat (rice). {0.67} ⇔ **Ami** bhAta khA.N . {1}
   (c) $\langle s_2, t_2 \rangle$: **They** eat (rice). {0.67} ⇔ **teo.N loke** bhAta **khAi** . {3}

The paradigm above illustrate that Assamese (and Bangla) verbs attest regular and consistent agreement properties with regard to person (but not with number and gender). Thus, the difference between (1a) and (1c), is such that modifying the target equivalent of *they* and changing it to *teo.N loke* is not going to resolve the translation conundrum here. The reason for this lies in the additional requirement of a change in the person agreement of the verb form. Altogether, (1b) and (1c) has the same FMS with the input s' but requires 1 and 3 edits respectively in order to obtain the translation of s'. We observe similar scenario for the English–Bangla translation example below.

2. (a) s': *I saw the boy.* ⇒ *Ami CheleTAke dekheChilAma.*
   (b) $\langle s_1, t_1 \rangle$: I saw the **girl**. {0.75} ⇔ Ami **meYeTAke** dekheChilAma. {1}
   (c) $\langle s_2, t_2 \rangle$: **He** saw the boy. {0.75} ⇔ **se** CheleTAke **dekheChila**. {2}

Both (2b) and (2c) have the same FMS score, but human post-editors need more effort if the TM-engine selects (2b) as the closest match instead of (2a) to obtain the translation of s'. Thus, even when two source-side sentences have the same FMS, their translation varies due to some some underlying property of the target-language which is not captured in standard FMS score.

[4] All example contains the input sentence to be translated s', the two closest matching sentence pair $\langle s_1, t_1 \rangle$ and $\langle s_2, t_2 \rangle$. The numbers in curly braces indicate FMS in the source-side and edit distance is the target-side.

4.1.2 Oblique Case Assignment

As discussed in 4.1, oblique case is implemented in English with a preposition. Hence, when an indirect object of a sentence receives oblique case, the preposition *to* surfaces as in (3a) and (3b). Example (3) illustrates the effect of oblique case assignment, when translated to Assamese.

3. (a) s': *I gave you a book.* $\Rightarrow$ *ma;i tomAka ekhana kitApa dilo.*
   (b) $\langle s_1, t_1 \rangle$: I gave a book **to you**.$\{0.5\}$ $\Leftrightarrow$ ma;i tomAka ekhana kitApa dilo. $\{0\}$
   (c) $\langle s_2, t_2 \rangle$: I gave **her** a book. $\{0.8\}$ $\Leftrightarrow$ ma;i **tAika** ekhana kitApa dilo. $\{1\}$

The above examples show that oblique case assignment is not surface apparent in Assamese and therefore their edges will mismatch in the translated output. The predicted output will look for changes in the target language with reference to the equivalent number of differences in the source language. However, the target language differs from the source language as no edits are required for these two sentences. The above example is more interesting in terms of TM match. The example pair in (3c) has much higher FMS (0.8) compared to the example pair in (3b) for the input in (3a). However, in order to obtain the translation of (3a), we require 1 edits from the target-side in (3c) while no change is require while considering (3b) as the TM match.

4.1.3 Classifier and Noun Concord

Often English articles are equated to the demonstrative in Bangla. Bangla demonstrative consists of classifiers e.g. *-TA*, *-khana*, *-jon*, etc. This property complicates FMS-based human post editing. This is shown in the subsentential segments in example (4).

4. (a) s': *a tiger* $\Rightarrow$ *ekTA bAgha*
   (b) $\langle s_1, t_1 \rangle$: a **box**$\{0.5\}$ $\Leftrightarrow$ ekaTA **bAksa**$\{1\}$
   (c) $\langle s_2, t_2 \rangle$: a **teacher** $\{0.5\}$ $\Leftrightarrow$ **ekajana shikShaka**$\{\mathbf{2}\}$

In above examples, we find the English article *a* takes different demonstrative in Bangla. The closest match in (4b) takes the same non-human demonstrative (*-TA*), similar to the source phrase in (4a). However, in (4c), *shikShaka* affix a different classifier *-jana*. Thus, the selection of (5c) as closest match will results more number of edits by the post editor. Thus, human translators may need to change the classifier depending on the property of the noun attached with it which is not reflected in the source-side (English) FMS.

4.2 Negation

In Assamese two sentential negative markers do not occur in the same sentence but a negative indefinite and the sentential negative element */nai/* can occur together. Like many other languages, in Assamese also negative polarity items and negative indefinites have the same shape. This is unlike English where the negative polarity items like anywhere, anything etc. are different from the negative indefinites. Therefore, in Assamese, the expressions for English anything anywhere is same as nothing, nowhere etc.

5. (a) s': *He forgave his brother.* $\Rightarrow$ *si tAra bhAYekaka mApha karile .*
   (b) $\langle s_1, t_1 \rangle$: He **never** forgave his brother. $\{0.8\}$ $\Leftrightarrow$ si tAra bhAYekaka **katiYAo** mApha **nakarile** . $\{2\}$
   (c) $\langle s_2, t_2 \rangle$: He forgave his **sister**. $\{0.75\}$ $\Leftrightarrow$ si tAra **bhanIYekaka** mApha karile . $\{1\}$

6. (a) s': *Nobody ate everything.* $\Rightarrow$ *koneo saba khowA nAi .*
   (b) $\langle s_1, t_1 \rangle$: Nobody ate **anything**. $\{0.67\}$ $\Leftrightarrow$ koneo **eko** khowA nAi .$\{1\}$
   (c) $\langle s_2, t_2 \rangle$: **He** ate everything. $\{0.67\}$ $\Leftrightarrow$ **si saba khAle .** $\{4\}$

Source and target language pairs indicate that additional negative elements are required in the target language output pairs. In both (5b) and (6b) the number of negative markers exceed the negative elements predicted from its equivalent source sentence.

Furthermore, in example (6), both (6b) and (6c) have same FMS (0.67) with (6a) but needs different human post-editing effort (1 and 4 respectively) to obtain the translation of (6a).

4.3 Translation Complexity of Verbal Features in Source and Target Languages

In this section, we focus on the effect of verbal features in TM-based human translation.

4.3.1 Copula

English attests the presence of a copula verb linking the subject and the object. However, many Indian languages delete the copula in similar constructions.

7. (a) s': *He is a secretary.* $\Rightarrow$ *tekheta ejana chekreterI .*
   (b) $\langle s_1, t_1 \rangle$: He **has** a secretary. $\{0.75\}$ $\Leftrightarrow$ **tekhetara** ejana chekreterI **AChe** .$\{2\}$
   (c) $\langle s_2, t_2 \rangle$: He is a **professor**. $\{0.75\}$ $\Leftrightarrow$ tekheta ejana **adhyApaka** . $\{1\}$

The source language pair above show that the copula verb *is* in (7a) indicates a person's occupation as a *secretary* whereas in the second sentence in (7b) the verb *has* leads to the meaning that a person employs a *secretary*. Translation of this pair leads to mismatched edges in the target language output as the copula verb does not appear in the translated output of (7a) whereas the verb indicating possession appears in the translation of (7b). Further mismatch is also created by the possessive case affix on the subject. In such cases, while finding a TM-match, the match in copula indicates lesser number of required human edits as reflected in example (7c). We find similar observations in English-to-Bangla translation direction.

8. (a) s': *The secretary is my father.* ⇒ *sekretarI holen amAra bAbA.*
   (b) $\langle s_1, t_1 \rangle$: The secretary is **your mother** . $\{0.60\}$ ⇔ sekretarI holen **tomAra mA** . $\{2\}$
   (c) $\langle s_2, t_2 \rangle$: The secretary is **a Hindu** . $\{0.60\}$ ⇔ sekretarI **ekjana hindu** . $\{3\}$

We found the copula verb (*is*) matches in examples (8a) and (8b) but does not match in (8c). The verb *is* in (8c) is not a copula verb. This effects the required translation edits, 3 and 2 edits for choosing (8b) and (8c) as TM-match respectively.

4.3.2 Complex Predicates

Preliminary investigations show that the complex verb in Assamese may bear resemblances to similar constructions in Hindi and Bengali. South Asian languages demonstrate the presence of aspectual complex predicates (Hook, 1974; Butt, 1995; Masica, 1991) where the complex predicates in question consist of two verbs: a main verb and a light verb. The main verb is in nonfinite form and bears lexical content, whereas the light verb bears tense and agreement features and other semantic features (Basu and Wilbur, 2010).

In Assamese also, in complex predicates the VV structure is such that the main verb appears first in the linear order and the two together behave like a single lexical verb with respect to diagnostics like scrambling and reduplication (see also (Butt, 1995) and (Ramchand, 2008)). There has been some analysis of Bangla Complex verbs in the theoretical framework of event semantics (Basu and Wilbur, 2010). In VC structures, two verbs occur together without any intervening material and their meaning and syntactic structures are always understood together. In Event Semantics, it can be explained as a type of construction where the two verbs actually represent a single event.

The semantic content of the second verb disappears and it only plays a role in attributing aspectual meaning. According to this framework, VCs are analysed as a *single* event having overtly grammaticalized internal sub-events (Basu and Wilbur, 2010). In the sentences below, the presence of a complex predicate in the translated output in (9b) lends a meaning which is not exactly compositional and also involves more lexical material than the input sentence in (9a). This is not the case when choosing (9c) as the closest possible match. Therefore the number of edits are more in (9b) than the closest match in (9c).

9. (a) s': *Ramesh will beat the dog.* ⇒ *rameshe kukurato mAriba .*
   (b) $\langle s_1, t_1 \rangle$: Ramesh will **kill** the dog. $\{0.8\}$⇔ rameshe kukurato **mAri pelAba** .$\{2\}$
   (c) $\langle s_2, t_2 \rangle$: Ramesh will beat the **man**. $\{0.8\}$ ⇔ rameshe **mAnuhajana** mAriba . $\{1\}$

4.3.3 Anaphora Resolution

Generative linguistic theory has shown that anaphoras across languages show different attributes with regard to their binding properties. Syntactic structures such as the following in (10a) and (10b) demonstrate certain complexities with regard to the problem of whether the proper noun and the pronoun are referring to the same individual.

10. (a) s': *Ramesh left after he found the bicycle* ⇒ *chAikelakhana powAra pAChata ramesha guchi ga'la .*
    (b) $\langle s_1, t_1 \rangle$: He left after Ramesh found the bicycle. $\{0.71\}$ ⇔ **ramesha** chAikelakhana powAra pAChata guchi ga'la . $\{2\}$
    (c) $\langle s_2, t_2 \rangle$: After he found the bicycle Ramesh left. $\{0.43\}$ ⇔ chAikelakhana powAra pAChata ramesha guchi ga'la .$\{0\}$

In the English sentences, in (10a) *Ramesh* and *he* can be understood as referring to the same person. This is in contrast to (10b) where *he* and *Ramesh* can be understood to be different people. The order of the two Noun Phrases is also significant. Linear order is shown not to be determinant of whether the pronoun is bound to the proper noun in (10c). However the translated counterparts show that anaphora resolution in Assamese does not work in the same way as English. In order for the pronoun and the proper noun to refer to the same individual, either the pronoun does not appear or the pronoun has to follow the proper noun immediately. The example in (10c) has much lower FMS compared to the example in (10b), but no edit is required while choosing the target t_2 for human post-editing to obtain the translation of the example in (10a).

5 Computational Treatment

In the previous section, we have seen examples where standard FMS produces an inadequate match for a particular input. Furthermore, we have listed some linguistic parameters which instantiate the problem of FMS-based TM match in a particular translation direction. The standard FMS computation does not encode any linguistic parameter for finding the closest possible TM-match. In this section, we propose some possible way of capturing the aforementioned linguistic parameters within the FMS computation to find better candidate matches, especially in the presence of ambiguity in the top rank FMS (c.f. Figure 1). These linguistic parameters vary from language to language. Here, we propose some possible solution towards English-to-Assamese and English-to-Bangla translation.

Handling Case and Agreement: In order to incorporate subject-verb agreement in the FMS score, we need to identify the source-language (English) verbs and its subject NP. This can be done using any parser available for the source-language. Both in Assamese and Bangla, the verb form changes with the PNG-features (derived using a source-language morphological analyzer) of the subject. In the case of TM-match, the mismatch in the subject position with different PNG-features leads to different verb inflections in the target-language (as in example (1b)). In such cases, a penalty needs to be introduced in the fuzzy match score by increasing the edit distance substitution cost (instead of uniform cost to all edit operations). Thus, example (1b) will have higher FMS compared to (1c).

Classifier concordance can be solved by looking into the feature structure of the source language noun attached to it. If the feature structures of the attached nouns differ in such a way that they take different classifiers (as in example (4)) then we can penalize the FMS score due to feature structure mismatch by increasing the cost of the edit operations. Thus, in example (4), the mismatch in *non-human* property of the words *tiger* and *teacher* will be penalized and will receive higher edit distance score and subsequently will produce low FMS between (4a) and (4c).

Handling Negation: Negation can be identified using a POS tagger of the language. In section 4.2, we have seen how negation can affect the FMS score-based TM match for human post editing (c.f example (5)). This can be easily solved by using POS tag information in the FMS score. If the differing item has a POS tag RB and the words are like not, n't, never, nothing etc. (Penn tagset[5] for English) then an additional penalty can be introduced in the FMS score. This is due to the fact that often a mismatch in a negative polarity word requires more edits in the target language (c.f. section 4.2). The introduction of penalty (higher cost for edit operation in Lavenshtein distance when any of the edit operation involves a NEG).

Handling Verbal Features: In the previous section, we have seen that certain verbal features introduce some complexity in translation which can not be captured using traditional FMS. However, some of these verbal features can be identified through syntactic or semantic processing. For example, copula verbs can be identified based on the syntactic property of the *be* verb which is often used to link the subject and the predicate of a sentence. As we have seen in example (7), a mismatch in the copula verb may need more attention during human post-editing in a CAT system, so we need to put some penalty for this situation.

Handling complex predicates in FMS computation is difficult as it depends on both source- and target-language sentence. Same source-language verb may behave as a complex predicate in some translation and may behave as a normal verb in some other cases. This is a target language phenomena during the process of English-to-Assamese and English-to-Bangla translation. In contrast, the FMS value is computed in the source side. In South Asian languages complex predicates have a fixed meaning (lexicalized), therefore we need listed entries to handle them in translation from a source language. For instance, if we take the root verb *mAra* and its suffixed form *mAri*, it can lead to different forms as a result of addition to other verbs. e.g. *mAri thak* (keep beating) *mAri de* (beat), *mAri pelA* (kill). Among all these verb-verb combinations only *mAri pelA* means something which is not *beat*. In all the other cases, the

[5] `http://www.mozart-oz.org/mogul/doc/lager/brill-tagger/penn.html`

verbal combinations are related to *beat*. If we find a mismatch in a verb from the potential complex predicate list, we will introduce some penalty due to their translation difficulty in the target language.

Identification of anaphora may help to disambiguate some pronouns in a sentence. The disambiguation of anaphora may help to understand the similarity of meaning between two sentences (as in example (10)).

6 Conclusion and Outlook

We have shown difficulties of estimating translation edits in a CAT based system using tradition FMS score. Traditional FMS score may end up with some spurious match in terms of number of edits required in target-side even when some other better candidate pair exists in the TM database. We have shown the misleading behaviour of traditional FMS in case of English-to-Assamese and English-to-Bangla translation directions. The similarity function can be further improved by encoding some morpho-syntactic information during FMS computation. This work can be further extended to provide a generic formula to encode morpho-syntactic information into the FMS computation.

References

D. Basu and R. Wilbur. 2010. Complex predicate in Bangla: An event-based analysis *Rice Working Papers in Linguistics*, **2**:1–19.

L. Bowker. 2002. Computer-aided translation technology: a practical introduction Chapter *Translation Memory Systems*, University of Ottawa Press, Ottawa, pp. 92-127.

M. Butt. 1995. The Structure of Complex Predicates in Hindi-Urdu Stanford, CA: CSLI Publications.

A. Chopde. 2001. ITRANS version 5.3 `http://www.aczoom.com/itrans/`

S. Dandapat and S. Morriessy and A. Way and J. van Genabith. 2012. CombiningEBMT, SMT, TM and IR Technologies for Quality and Scale In *Proceedings of the EACL 2012 Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, pages 48–58, Avignon, France.

A. Dara and S. Dandapat and D. Groves and J. van Genabith. 2013. TMTprime: A Recommender System for MT and TM Integration In *the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies , NAACL HLT 2013*, Atlanta, GA.

M. Espla and F Sanchez-Martinez and M. L. Forcada. 2011. Using Word Alignments to Assist Computer-aided Translation Users By Marking Which Target-side Words to Change or Keep Unedited In *Proceedings of the 15th Annual Meeting of the European Association for Machine Translation (EAMT 2011)*, pages 81–88, Leuven, Belgium.

M. Espla and F Sanchez-Martinez and M. L. Forcada. 2011. Target-Language Edit Hints: a Basic Description of the Method. In Technical Report. Dep. de Llenguatgesi Sistemes Inform'atics, Universitat d'Alacant, Spain.

Y. He and Y. Ma and J. vanGenabith and A. Way. 2010. Bridging SMT and TM with Translation Recommendation In *Proceedings of the 48th Annual Meeting of the Association of Computational Linguistics (ACL 2010)*, page 622630, Uppsala, Sweden.

P. E. Hook. 1974. The Compound Verb in Hindi University of Michigan, Center for South and Southeast Asian Studies.

M. Kay. 1980. The Proper Place of Men and Machines in Language Translation Technical report, CSL-80-11, Xerox Palo Alto Research Center, Palo Alto, Calif. Reprinted in *Machine Translation* 1997, **12**:3–23.

P. Koehn and J. Senellart. 2010a. Fast Approximate String Matching with Suffix Arrays and A* Parsing In *Proceedings of the 9th Annual Conference of the Association for Machine Translation in Americas (AMTA 2010)*, page 4557, Denver, CO.

P. Koehn and J. Senellart. 2010b. Convergence of Translation Memory and Statistical Machine Translation In *Proceedings of the e 2nd Joint EM+/CNGL, Workshop "Bringing MT to the User: Research on Integrating MT in the Translation Industry"*, page 2131, Denver, CO.

J. Laveling and D. Ganguly and S. Dandapat and G. F. Jones. 2012. Approximate Sentence Retrieval for Scalable and Efficient Example-based Machine Translation In *Proceedings of the 24th International Conference on Computation Linguistics (COLING 2012)*, pages 15711586, Mumbai, India.

V. I. Levenshtein. 1965. inary Codes Capable of Correcting Deletions, Insertions, and Reversals *Doklady Akademii Nauk SSSR*, **163(4)**:845–848.

C. P. Masica. 1991. The Indo-Aryan Languages Cambridge University Press.

G. Ramchand. 2002. Verb Meaning and the Lexicon Cambridge University Press.

R. Sikes. 2007. Fuzzy Matching in Theory and Practice *Multilingual*, **18(6)**:39–43.

K. Wang and C. Zong. and K Su. 2013. Integrating Translation Memory into Phrase-Based Machine Translation during Decoding In *Proceedings of the 51st Annual Meeting of the Association of Computational Linguistics (ACL 2013)*, page 1121, Sofia, Bulgaria.

V. Zhechev and J. Genabith 2010. eeding Statistical Machine Translation with Translation Memory Output through Tree-based Structural Alignment In *Proceedings of the COLING'10, Workshop on Syntax and Structure in Statistical Translation*, page 4351, Beijing, China.