A Framework for Example-Based Translation Aid Tools

Stelios Piperidis, Harris Papageorgiou, Iason Demiros,

Christos Malavazos, Yiannis Triantafyllou

Institute for Language and Speech Processing
Artemidos & Epidavrou, 151 25 Marousi
{spip, xaris, iason, christos, yiannis} @ilsp.gr

Abstract

This paper describes an approach to develop a computer-aided translation (CAT) platform dedicated to the derivation of translation examples in an attempt to maximise past translation reuse as well as terminology and style consistency in the translation of specific types of text. Such tools have come in the bibliography under the term Translation Memory (TM) tools. We first briefly outline the setting in which the development effort was undertaken, and then we give a description of the architecture of the platform. Current applications building on the technology underlying CAT and translation example derivation are also briefly discussed in the context of multilingual information retrieval.

Keywords : Machine Translation, Example-Based Machine Translation, Translation Memory, Alignment, Text Matching, Best Match Retrieval.

1. Introduction

Researchers have been working on machine translation of natural languages for almost 50 years. The reasons for this have varied from R&D interest to commercial payoff. While there have been some successful attempts and a few commercial systems, high quality, fully automatic machine translation remains an elusive goal. Not surprisingly, there is some disagreement about how best to proceed. On one side, researchers working on knowledge-based approaches argue that to obtain high quality translation requires considerable linguistic knowledge and large knowledge bases. On the other side, researchers advocating statistical approaches argue that it is impractical to build large enough knowledge bases to make this feasible, but large corpora of translated text do exist that can be used to train a statistics-based system. There has been a lot of rhetoric on both sides but neither approach has made much progress. But either way, the statistical approach is producing a set of useful terminology and reuse tools. Unlike traditional Machine Translation (MT) with its largely

automatic approaches, these tools do not attempt to compete with the human at what the human does best (translating the easy vocabulary and the easy grammar), but complement the human in areas where they know they need help (difficult vocabulary and reuse).

For the future, we can expect non-toy MT systems to be hybrid. Symbolic approaches function better on phenomena exhibiting regular linguistic behaviour while statistical approaches handle phenomena with little regular behaviour. Hybrid cooperative methods seem to be the only way forward in MT.

2. Background

2.1 Translation Memory

The purpose of our work has been to develop a translation aid tool dedicated to managing repetition phenomena in the translation of specific types of text. Such tools have come in the bibliography under the term Translation Memory (TM) tools.

The idea behind Translation Memories is that in the process of translation, blocks of text, ranging from simple sentences to whole chapters, can be reused. This becomes more evident in particular types of text such as legal documents (contracts, regulations, etc.), technical documentation texts (manuals, etc.) but also in almost all sorts of texts bearing formulaic flow and structure. Appropriate storage of pairs of source language (SL) and target language (TL) blocks of text and provision of means for retrieval of applicable solutions and post-editing them would increase the productivity of a translator and at the same time improve the quality and consistency of the translation.

2.2 Technology behind Translation Memories

The technology underlying translation memory applications stems from what has been described in the literature as example-based machine translation (EBMT). EBMT is based on the idea of performing translation by imitating translation examples of similar sentences (Nagao 1994). In this type of translation system, a large amount of bi/multi-lingual translation examples has been stored in a textual database and input expressions are rendered in the target language by retrieving from the database that example which is most similar to the input.

There are three key issues which pertain to example-based translation (Piperidis 1995):

- establishment of correspondence between units in a bi/multi-lingual text at sentence, phrase or word level, i.e. alignment of parallel texts
- a mechanism for retrieving from the database the unit that best matches the input
- exploitation of the retrieved translation example to produce the actual translation of the input sentence

2.3 Correspondence between text units

Several different approaches have been proposed tackling the alignment problem at various levels. Catizone's technique (Catizone, Russell & Warwick 1989) was to link regions of text according to the regularity of word cooccurrences across texts. Brown described a method based on the number of words that sentences contain

(Brown 1991). Moreover, certain anchor points and paragraph markers are also considered. The method has been applied to the Hansard Corpus and has achieved an accuracy between 96%-97%.

Gale proposed a method that relies on a simple statistical model of character lengths (Gale 1991). The model is based on the observation that longer sentences in one language tend to be translated into longer sentences in the other language while shorter ones tend to be translated into shorter ones. Although the apparent efficacy of the Gale-Church algorithm is undeniable and validated on different pairs of languages (English - German - French - Czech - Italian), it seems to be awkward when handling complex alignments. Complex alignments are defined to be alignments in which the 1-1 correspondence between text units in the parallel texts does not hold, and they are usually due to mergers of sentences occurring during the translation process.

To overcome the inherited weaknesses of the Gale-Church method, (Simard 92) proposed using cognates, which are pairs of tokens of different languages which share "obvious" phonological or orthographic and semantic properties, since these are likely to be used as mutual translations.

2.4 Best Match retrieval

In establishing a mechanism for the best match retrieval two crucial tasks are identified:

- determining whether the search is for matches at sentence or sub-sentence level, that is determining the "text unit", and
- the definition of the metric of similarity between two text units.

As far as the decision about the text unit is concerned, the obvious choice is to use as text unit the sentence. This is because, not only are sentence boundaries mostly unambiguous, but also translation proposals at sentence level is what a translator is usually looking for. Sentences can, however, be quite long. And the longer they are, the less possible it is that they will have a perfect match in the translation archive, and the less flexible the EBMT system will be.

On the other hand, if the text unit is the sub-sentence, we face one major problem, that is the possibility that the resulting translation of the whole sentence will be of low quality, due to boundary friction and incorrect chunking. In practice, EBMT systems that operate at sub-sentence level involve the dynamic derivation of the optimum length of segments of the input sentence by analysing the available parallel corpora. This requires a procedure for determining the best "cover" of an input text by segments of sentences contained in the database (Nirenburg 1993). It is assumed that the translation of the segments of the database that cover the input sentence is known. What is needed, therefore, is a procedure for aligning parallel texts at sub-sentence level (Kaji, Kita & Morimito 1992), (Sadler & Vendelmans 1990). If sub-sentence alignment is available, the approach is fully automated but is quite vulnerable to the problem of low quality as well as to ambiguity problems when the produced segments are rather small. Despite the fact that almost all running EBMT systems employ the sentence as the text unit, it is believed that the potential of EBMT lies on the exploitation of fragments of text smaller than sentences and the combination of such fragments to produce the translation of

whole sentences (Sato & Nagao 1990). Automatic sub-sentential alignment is, however, a problem yet to be solved.

Turning to the definition of the metric of similarity, the requirement is usually twofold. The similarity metric applied to two sentences (by sentence we refer to both sentence and sub-sentence fragment) should indicate how similar the compared sentences are, and perhaps the parts of the two sentences that contributed to the similarity score. The latter could be just a useful indication to the translator using the EBMT system, or a crucial functional factor of the system as will be later explained.

The similarity metrics reported in the literature can be characterised depending on the text patterns they are applied on. Thus, the word-based metrics compare individual words of the two sentences in terms of their morphological paradigms, synonyms, hyperonyms, hyponyms, antonyms, pos tags (Nirenburg 1993) or use a semantic distance d ($0 \leq d \leq 1$) which is determined by the Most Specific Common Abstraction (MSCA) obtained from a thesaurus abstraction hierarchy (Sumita & Iida 1991). Then, a similarity metric is devised, which reflects the similarity of two sentences, by combining the individual contributions towards similarity stemming from word comparisons.

The word-based metrics are the most popular, but other approaches include syntax-rule driven metrics (Sumita & Tsutsumi 1988), character-based metrics (Sato 1992) as well as some hybrids, (Cranias, Papageorgiou & Piperidis 1994), (Boutsis & Piperidis 1996). The character-based metric has been applied to Japanese, taking advantage of certain characteristics of Japanese. The syntax-rule driven metrics try to capture similarity of two sentences at the syntax level. This seems very promising, since similarity at the syntax level, perhaps coupled by lexical similarity in a hybrid configuration, would be the best an EBMT system could offer as a translation proposal. The real time feasibility of such a system is, however, questionable, since it involves the complex task of syntactic analysis.

2.5 Exploitation of the retrieved translation example

The third key issue of EBMT, that is exploiting the retrieved translation example, is usually dealt with by integrating into the system conventional MT techniques (Kaji, Kida & Morimoto 1992), (Sumita & Iida 1991). Simple modifications of the translation proposal, such as word substitution, would also be possible, provided that alignment of the translation archive at word level was available.

3 A Proposal for a new Alignment algorithm

In the above framework of EBMT a task of crucial importance is the establishment of correspondences between units of multilingual texts at sentence, phrase or even word level. The adopted criteria for ascertaining the adequacy of alignment methods are stated as follows :

- an alignment scheme must cope with the embedded extra-linguistic data (tables, anchor points, SGML markers, etc) and their possible inconsistencies.

- it should be able to process a large amount of texts in linear time and in a computationally effective way.
- in terms of performance a considerable success rate (above 99% at sentence level) must be encountered in order to construct a database with truthfully correspondent units. It is desirable that the alignment method is language-independent.
- the proposed method must be extensible to accommodate future improvements. In addition, any training or error correction mechanism should be reliable, fast and should not require vast amounts of data when switching from a pair of languages to another or dealing with different text type corpora.

In the next sections, we propose an alignment scheme in order to deal with the complexity of varying requirements envisaged by different applications in a systematic way (Papageorgiou 96),(Papageorgiou 97). Our approach is based on several observations. First of all, we assume that establishment of correspondences between units can be applied at sentence, clause, and phrase level. Alignment at any of these levels has to invoke a different set of textual and linguistic information (acting as unit delimiters). In 3.1 we give briefly the details of the algorithm and in 3.2 we discuss alignment evaluation at sentence level.

3.1 Alignment Algorithm

Content words, unlike functional ones, might be interpreted as the bearers that convey information by denoting the entities and their relationships in the world. The notion of spreading the semantic load supports the idea that every content word should be represented as the union of all the parts of speech we can assign to it (Basili 92). The postulated assumption is that a connection between two units of text is established if, and only if, the semantic load in one unit approximates the semantic load of the other.

Based on the fact that the principal requirement in any translation exercise is meaning preservation across the languages of the translation pair, we define the semantic load of a sentence as the patterns of tags of its content words. Content words are taken to be verbs, nouns, adjectives and adverbs. The complexity of transfer in translation imposes the consideration of the number of content tags which appear in a tag pattern. By considering the total number of content tags the morphological derivation procedures observed across languages, e.g. the transfer of a verb into a verb+deverbal noun pattern, are taken into account. Morphological ambiguity problems pertaining to content words are treated by constructing ambiguity classes (acs) leading to a generalised set of content tags (Papageorgiou et al. 94).

It is essential here to clarify that in this approach no disambiguation module is prerequisite. The time breakdown for morphological tagging, without a disambiguator device, is according to (Cutting 92) in the order of 1000 ìseconds per token. Thus, tens of megabytes of text may then be tagged per hour and high coverage can be obtained without prohibitive effort.

Having identified the semantic load of a sentence , Multiple Linear Regression is used to build a quantitative model relating the content tags of the source language (SL) sentence to the response, which is assumed to be the sum of the counts of the corresponding content tags in the target language (TL) sentence. The regression model is fit to a set of sample data which has been manually aligned at sentence level. Since we

intuitively believe that a simple summation over the SL content tag counts would be a rather good estimator of the response, we decide that the use of a linear model would be a cost-effective solution.

The linear dependency of y (the sum of the counts of the content tags in the TL sentence) upon x_i (the counts of each content tag category and of each ambiguity class over the SL sentence) can be stated as :

$$y=b_0+b_1x_1+b_2x_2+b_3x_3+...+b_nx_n+\varepsilon \quad (1)$$

where the unknown parameters $\{b_i\}$ are the regression coefficients, and ε is the error of estimation assumed to be normally distributed with zero mean and variance σ^2.

In order to deal with different taggers and alternative tagsets, other configurations of (1), merging acs appropriately, are also recommended. For example, if an acs accounts for unknown words, we can use the fact that most unknown words are nouns or proper nouns and merge this category with nouns. We can also merge acs that are represented with only a few distinct words in the training corpus. Moreover, the use of relatively few acs (associated with content words) reduces the number of parameters to be estimated, affecting the size of the sample and the time required for training.

The method of least squares is used to estimate the regression coefficients in (1). Having estimated the b_i and σ^2, the probabilistic score assigned to the comparison of two sentences across languages is just the area under the $N(0,\sigma^2)$ p.d.f., specified by the estimation error. This probabilistic score is utilised in a Dynamic Programming (DP) framework similar to the one described in (Gale 91). The DP algorithm is applied to aligned paragraphs and produces the optimum alignment of sentences within the paragraphs.

3.2 Alignment Evaluation

The application on which we are developing and testing the method is implemented on the Greek-English language pair of sentences of the CELEX corpus (the computerised documentation system on European Community Law).

Training was performed on 40 Articles of the CELEX corpus accounting for 30000 words. We have tested this algorithm on a randomly selected corpus of the same text type of about 3200 sentences. Due to the sparseness of acs (associated only with content words) in our training data, we reconstruct (1) by using four variables. For inflective languages like Greek, morphological information associated to word forms plays a crucial role in assigning a single category. Moreover, by counting instances of acs in the training corpus, we observed that words that, for example, can be a noun or a verb, are (due to the lack of the second singular person in the corpus) exclusively nouns. Hence :

$$y=b_0+b_1x_1+b_2x_2+b_3x_3+b_4x_4+\varepsilon \quad (2)$$

where x_1 represents verbs, x_2 stands for nouns, unknown words, vernou (verb or noun) and nouadj (noun or adjective), x_3 adjectives and veradj (verb or adjective), x_4 adverbs and advadj (adverb or adjective). σ^2 was estimated at 3.25 on our training sample, while the regression coefficients were:

$$\mathbf{b_0 = 0.2848, b_1 = 1.1075, b_2 = 0.9474, b_3 = 0.8584, b_4 = 0.7579}$$

An accuracy that approximated a 100% success rate was recorded. Results are shown in Table 1. It is remarkable that there is no need for any lexical constraints or certain anchor points to improve the performance. Additionally, the same model and parameters can be used in order to cope with the infra-sentence alignment.

Category	N	correct matches
1-0 or 0-1	5	4
1-1	3178	3178
2-1 or 1-2	36	35
2-2	0	0

Table 1 : Matches in sentence pairs of the CELEX corpus

4. Components of a Corpus-based Translation Drafting Tool

4.1 Translation Memory building blocks

The key issues in the process of building a translation memory tool revolve around several major axes :

- organisation of multilingual parallel corpora, i.e. texts in different languages, one being the translation of the other
- alignment of parallel texts, i.e. establishment of correspondences between units of parallel texts
- textual database management facilities
- text matching techniques
- linguistic processing tools at the monolingual level and its role in text matching
- powerful known-how-to-use editing environments

The basic building blocks of a typical TM system are :

- the text handling tool

- the alignment tool
- the database storage tool
- the text matching tool
- the graphical user interface
- the text lemmatisation and tagging tools (if available)

4.2 Text handling

Text handling can, in general, be seen as a sophisticated interface between input text streams and various text processing modules. At the stage of analysis, the text handler has the responsibility of transforming a text from the original form in which it is found into a form suitable for the manipulation required by the application; at the stage of synthesis, it is responsible for the reverse process, i.e. for converting the output text from the form used by the application into a form equivalent to that of the input text. The text handler has been used only during two separate procedures:

During the phase of translation memory building: before loading the existing corpus to the database, it serves as the interface between the normalisation and alignment procedures, marking in each text the structural units and the extra-linguistic elements

During the translation phase, the text handler is used before the sentence matching algorithm in order to detect sentence boundaries and extra-linguistic elements included in the text to be translated.

4.3 Text Alignment

Alignment consists in establishing correspondence links between units in a bi/multi-lingual text. The level at which alignment is performed in our TM is that of the sentence. The heart of the alignment scheme is proposed to be a method for aligning sentences based on a simple statistical model of character lengths. The method relies on the assumption that longer sentences in the source language tend to be translated into longer sentences in the target and that shorter sentences in the source are translated into shorter sentences in the target. A probabilistic score is assigned to each pair of proposed sentence pairs, based on the ratio of lengths of the sentences and the variance of this ratio. This probabilistic score is used in a dynamic programming framework in order to find the maximum likelihood alignment of sentences. The whole process proceeds in two steps. First, paragraphs are aligned and then sentences within a paragraph are aligned. The aligner module takes as input handled texts, thus exploiting sentence and paragraph markers. Text alignment is performed only at the text corpora pre-processing, i.e. data or meta-data preparation stage. It is essentially a process of building a multilingual lexicon at the level of sentences. Text alignment data are then stored away in the translation memory (reference corpora) database.

4.4 Text linguistic annotation

In case the necessary linguistic tools are in place, parallel corpora are linguistically annotated. Lemmatisation and morphological annotation, mainly Part-Of-Speech (POS) tagging are the minimum processing procedures, the results of which are stored in the translation memory database. Lemmatisation

consists in deriving the lemma or canonical form of each wordform while tagging consists in labelling each wordform, with its grammatical category or part of speech.

For the purpose of considering linguistic characteristics during a matching procedure, lemmas and POS tags are the main entities to take into account.

In addition to morphological annotation of the words of a sentence, further linguistic information is recommended to be coded. This includes foreign words, letters and numbers, abbreviations, dates, etc.

4.5 Translation Memory Database

The complexity inherent in the translation processes within the basic mechanisms of a typical TM system, necessitates the existence of well-defined powerful resources. Apart from that, the need for real-time responses of a TM system brings up the subject of an efficient as well as a complete database structure that should meet the following requirements:

- a friendly and non-complicated database-interface.
- portability of the database information
- system independent of DBMS platform (ability to work with databases of several types such as Jet Engine, ISAM, SQL, ORACLE)
- Desktop intelligence.
- Sharing the resources most optimally.
- Optimal network utilization.
- Performance and System Management.
- Security and Access Control.

The database sessions during a translation phase are automatically controlled by the system which is entirely responsible for the communication with the underlying DBMS. Since the user requests are translated to database queries automatically, the user is not presented with any unnecessary technical details concerning the database procedures. Even in the case of a remote database, the system completely hides the underlying communication methodology from the user.

4.6 Text Matching

The corpus preprocessing and database storage stages render the corpus in the appropriate form for the application. The matching tool searches for database sentences of language A that are identical or only similar to an input sentence (in language A) and retrieve the equivalent sentence in language B. The approach adopted to text matching is based on computations of common elements between an input sentence and a database sentence and computation of consecutive elements in them. Computations are either based on wordforms and their respective position in the compared sentences or on lemma-tag tuples of each word in the compared sentences as well as their respective positions in them.

The matching mechanism consists of two processes:

(i) the perfect match process by which the system finds a database sentence (and its translation) in the TM which is identical to the input sentence, and

(ii) extraction of candidate sentences and the fuzzy match process. The fuzzy match process aims at extracting from the TM a sentence and its translation which resembles the given input sentence above a certain minimum degree (percentage).

4.7 Translation Memory enhancements

We are currently developing a number of supplementary modules to the TM such as:

- a terminological database to store terms and use them during the matching procedure
- an interactive alignment tool to detect and correct misalignments
- sub-sentence alignment (at phrase or word level)
- porting of the TM in other platforms

5 References

[Boutsis S. & Piperidis S. 96], "Automatic Extraction of Bilingual Lexical Equivalences from Parallel Corpora", *Proceedings of Mulsaic/ECAI 96, Budapest, Hungary,* 1996.

[Basili R. Pazienza M. Velardi P. 92] "Computational lexicons: The neat examples and the odd exemplars". Proc. of the Third Conference on Applied NLP 1992

[Brown P.F. et al 91], "Aligning Sentences in Parallel Corpora", *Proceedings of the 29th Annual Meeting of the ACL.,* 1991. pp.169-176.

[Catizone R., Russel G., Warwick S. 89], "Deriving translation data from bilingual texts", *Proceedings of the First Lexical Acquisition Acquisition Workshop,* Detroit, 1989.

[Cranias L., Papageorgiou H. and Piperidis S. 94], "A matching technique in Example-Based Machine Translation", *Proceedings of Coling,* 1994. pp.100-105

[Cutting D. Kupiec J. Pedersen J. Sibun P. 92] "A practical part-of-speech tagger " Proc.of ACL 1992

[Gale W.A. and Church K.W. 91], "A Program for Aligning Sentences in Parallel Corpora", *Proceedings of the 29th Annual Meeting of the ACL.,* 1991. pp.177-184.

[Kaji H., Kida Y. and Morimoto Y. 92], "Learning Translation Templates from Bilingual Text". *Proceedings Coling.,* 1992. pp.672-678.

[Nagao M. 94], "A framework of mechanical translation between Japanese and English by analogy principle". *Artificial and Human Intelligence, ed. Elithorn A. and Banerji R., North-Holland,* 1994. pp.173-180.

[Nirenburg S. et al 93], " Two Approaches to Matching in Example-Based Machine Translation", *Proceedings of TMI-93, Kyoto, Japan,* 1993.

[Papageorgiou H., Cranias L. and Piperidis S. 94], "Automatic Alignment in Parallel Corpora", *Proceedings of the 32nd Annual Meeting of the ACL,* 1994.

[Papageorgiou H. 96], "Hybrid Techniques in Processing of Parallel Bilingual Corpora" Ph.D. Thesis, 1996

[Papageorgiou H. 97] "Clause Recognition in the Framework of Alignment" in *Recent Advances in Natural Language Processing, ed. R. Mitkov & N. Nicolov, J.Benjamins, 1997, pp.417-426.*

[Piperidis S. 95], "Interactive corpus-based translation drafting tool (Translearn)", *Aslib Proceedings*, Vol.47, no.3, March 1995. pp.83-92.

[Sadler V. and Vendelmans R. 90], "Pilot Implementation of a Bilingual Knowledge Bank". *Proceedings of Coling,* 1990. pp.449-451.

[Sato S. and Nagao M. 90], "Towards Memory-based Translation", *Proceedings of Coling,* 1990. pp.247-252.

[Sato S. 92], "CTM: An Example-Based Translation Aid System", *Proceedings of Coling,* 1992. pp.1259-1263.

[Simard 92] Simard M. Foster G. Isabelle P. "Using cognates to align sentences in bilingual corpora" Proc. of TMI 1992

[Sumita E. and Iida H. 91], " Experiments and Prospects of Example-based Machine Translation", *Proceedings of the 29th Annual Meeting of the ACL.,* 1991. pp.185-192.

[Sumita E. and Tsutsumi Y. 88], "A Translation Aid System Using Flexible Text Retrieval Based on Syntax-Matching", *TRL Research Report, Tokyo Research Laboratory, IBM,* 1988.