Exploitation of Parallel Texts for Populating MT & TM Databases

Anna Samiotou, Lambros Kranias, George Papadopoulos, Marita Asunmaa, Gudrun Magnusdottir

ESTeam AB
Sikelianou 8, 146 71 Athens, Greece
esteam@otenet.gr

Abstract

Parallel texts are an important resource for applications in multilingual natural language processing and human language technology. This paper presents a method for exploiting available parallel texts, both human translated and revised machine translated texts in order to populate machine translation and translation memory databases.

1. Introduction

Parallel texts play an important role in Machine Translation (MT) and multilingual natural language processing. They are rich resources for development of monolingual, bilingual and multilingual resources both for new language pairs and for existing language pairs for a specific domain to be used in a number of natural language processing applications, for automatic lexical acquisition (e.g. Gale and Church, 1991; Melamed, 1997), etc.

This paper presents a method for populating MT and Translation Memory (TM) databases by exploiting parallel texts. The method deploys selected legacy data from the domain(s) under investigation and available parallel texts both human translated and revised machine translated texts. The software used is the ESTeam Translator© (ET) software[1] (ESTeam AB, 2004), a data-driven multilingual translation software product which integrates MT and TM technology to produce a full translation in one or multiple languages.

Current applications using the presented method include the creation of new LRs for the languages of the new members of the EU and their linking to all the existing EU languages as well as the creation of LRs for the translation needs of the Athens Organising Committee for the Olympic Games 2004.

The paper is organised as follows: Section 2 gives an overview of the methodological approach for processing parallel texts. Section 3 provides a brief outline for the application of the method in a commercial project. Finally, Section 4 concludes the paper.

2. Processing Parallel Texts

Translated data is a rich resource to solving translation problems. This resource has yet to be explored to its full extent (Isabelle et al., 1993). ESTeam applies pre-processing on a monolingual level as well as alignment in order to domain-tune lexical resources as well as extract translation equivalents on multiple levels.

2.1 Pre-processing

The monolingual data is structured into domains and analysed in three processing levels, that is, sentences, sub-sentences and words (tokenisation). The segmented text is sorted per level according to the frequency of occurrence and then, words and frequent collocations can be imported into the MT lexicon and sentences and sub-sentences into the TM database.

The processing on the monolingual level of the parallel texts is important since the monolingual data provides a resource for extensive multilingual linking. ET uses monolingual data to map to any other language once the data becomes available through resources or translation interaction.

Any general purpose lexical resource lacks information about domain. The information on the frequency of the units gives indications within the domains on which units have to be translated with priority (i.e. the high frequent ones). It also indicates which units are likely to be incorrect such as misspellings coming from wrong typing or scanning errors (i.e. the very low frequent ones) and this is judged on both frequency and similarity criteria. This information is used to automatically structure the lexical data for any domain when building the multilingual lexica (see example in *Figure 1*).

Language	*Unit*	*Domain*	*Frequency*
French	fils	Computer/Textiles	1011/741
English	threads	Computer/Textiles	14/573
English	yarns	Computer/Textiles	1/620
English	wires	Computer/Textiles	994/0

Figure 1. Example of Domain Tuning

In MT any monolingual data is deployed as a target language resource. When a source unit has multiple translations into another language, frequency information relating to the context of the target units gives indication on which translation alternatives MT automatically selects, i.e. the stronger the statistical indication is, the more likely it is to be selected (see example in *Figure 1*). ET also calculates the frequency of the translation links by combining source and target frequencies per domain (see example in *Figure 2*).

French⇔English	*Domain*	*Link Frequency*
fils ⇔ threads	Computer/Textiles	1025/1314
fils ⇔ yarns	Computer/Textiles	1012/1361
fils ⇔ wires	Computer/Textiles	2001/741

Figure 2. Example of Statistical Disambiguation

[1] http://www.esteam.gr

Context statistics are also calculated on the monolingual data, in order to assign weights on the co-occurrence of words and contribute to the word sense disambiguation within the same domain. In the examples in *Figures 1 & 2*, the English units *threads* & *yarns* win over *wires* as translations of the French unit *fils* in the Textiles domain. If the input French unit is: *fils de coton* and the context statistic model run on the monolingual data had calculated:

- *cotton threads* (100)
- *cotton yarns* (5)

then the *cotton threads* wins.

The more correct legacy monolingual data in the TM the better when using the ET, because it serves for target language verification (TLV), i.e. the machine translation result is automatically post-edited by the target language TM data (sentences and/or sub-sentences) based on a number of criteria permitting actions such as deletion or addition of functional units, changing word order and morphological variations. Example:

- input French source unit for translation:
  - *fils de coton*
- suggested translations in English:
  - *threads of cotton*
  - *yarns of cotton*
- existing units in the English TM:
  - *cotton threads*

=> TLV disambiguates and post-edits

2.2 Alignment

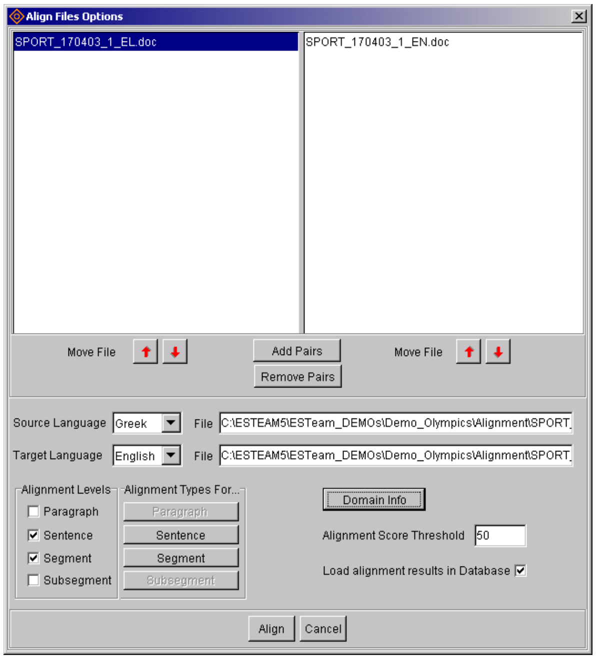
ET Aligner aligns parallel texts in different languages and at sentence and sub-sentence level (Kranias, 1995). The ET Aligner requires file, paragraph or sentence aligned parallel text. Assuming that a document is a hierarchical structure where the top level is the document itself and the deeper levels are paragraphs, sentences, sub-sentences and finally words, the ET Aligner, takes as input two parallel documents and automatically aligns them at the aforementioned deeper levels.

Alternatively, the ET Aligner processes pre-aligned documents at a given level and aligns them at a deeper level (e.g. if the given level is paragraph then it further processes at sentence and sub-sentence level). *Figure 3* displays the ET Aligner user interface. The supported format of the input documents is plain text in UTF-8 encoding, html, Microsoft Word document and TMX. The output results are in TMX format (see *Figure 4*)

At each level the ET Aligner uses a Dynamic Programming algorithm in order to detect the optimal text unit correspondences. The Aligner evaluates a number of criteria, mainly statistical and lexical information, in order to produce corresponding text unit pairs at each level, such as:

- the number of words per unit
- the number of characters per unit
- existence of strings such as numbers and dates
- special treatment of non-content words such as articles and prepositions
- special treatment of characters such as parenthesis and square brackets
- advanced lexicon look-up

The user can specify the criteria to be used and assign a weight to each criterion. Based on the previous, the Aligner assigns a reliability score to each produced aligned pair.



SPORT_170403_1_EL.doc

SPORT_170403_1_EN.doc

Source Language Greek File C:\ESTEAM5\ESTeam_DEMOs\Demo_Olympics\Alignment\SPORT_

Target Language English File C:\ESTEAM5\ESTeam_DEMOs\Demo_Olympics\Alignment\SPORT_

Alignment Levels

- □ Paragraph
- ☑ Sentence
- ☑ Segment
- □ Subsegment

Alignment Types For...

- Paragraph
- Sentence
- Segment
- Subsegment

Alignment Score Threshold 50

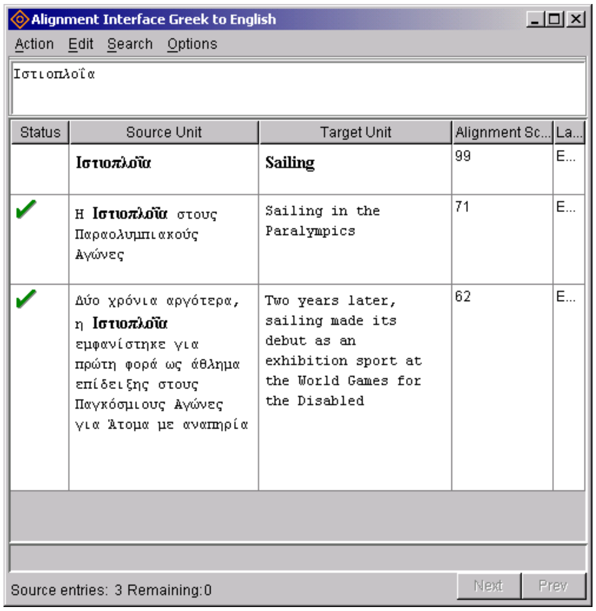
Load alignment results in Database ☑

Figure 3. ET Aligner User Interface

The alignment results are imported in a separate database, the ALIGN database. High quality alignment results are directly imported in the TM and/or MT databases. Medium and possibly low quality alignment results can be browsed and edited through the user-friendly ET Alignment Browser & Editor (see *Figure 5*) and the accepted and/or edited by the user results are imported in the TM.

```
<tu tuid="1-1" segtype="sentence">
    <prop type="x-ORGN">EL.doc_EN.doc.tmx</prop>
    <prop type="x-DOMN">0~Olympics~*/</prop>
    <prop type="x-ALGN">,01,,.</prop>
    <prop type="x-VALD">71</prop>
    <tuv lang="EL">
        <seg>Η Ιστιοπλοΐα στους Παραολυμπιακούς Αγώνες</seg>
    </tuv>
    <tuv lang="EN">
        <seg>Sailing in the Paralympics</seg>
    </tuv>
</tu>
```

Figure 4. Example of Alignment Results in TMX



Ιστιοπλοΐα

Status	Source Unit	Target Unit	Alignment Sc...	La...
	Ιστιοπλοΐα	**Sailing**	99	E...
✓	Η **Ιστιοπλοΐα** στους Παραολυμπιακούς Αγώνες	Sailing in the Paralympics	71	E...
✓	Δύο χρόνια αργότερα, η **Ιστιοπλοΐα** εμφανίστηκε για πρώτη φορά ως άθλημα επίδειξης στους Παγκόσμιους Αγώνες για Άτομα με αναπηρία	Two years later, sailing made its debut as an exhibition sport at the World Games for the Disabled	62	E...

Source entries: 3 Remaining:0

Figure 5. ALIGN Database Browser & Editor

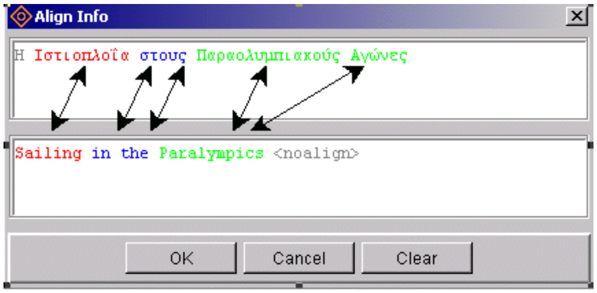
The ET Alignment Browser & Editor offers to the user full control over the alignment results which are stored in the ALIGN database. Its main features include:

- Various search modes (full/fuzzy match, word(s) in context, dynamic searches, combined source-target searches) for browsing the database contents
- Insert, Modify, Delete actions on selected contents
- Controlled global text replacements
- Logging of all user actions
- Dynamic Import/Export of the database contents to the TM

2.3 Word-Alignment Information

In the ET system, word-alignment information is available, through the alignment process (Meyers 1998, Ahrenberg et al, 2000) by the use of an MT lexicon of words and phrases. Word-alignment information defines the translation links between words of reference-SL and reference-TL text units (the TM pair), in other words it defines which word/phrase of the $S_{ref\text{-}SL}$ translates to which word/phrase of the $S_{ref\text{-}TL}$ (and can, in general, include phrases with non-consecutive words).

The MT lexicon defines the relevance of two text units being compared, by defining translation links between their words, and then puts a marker on the corresponding word-alignment information to be later used for the application of Fuzzy Match Post Editing (Kranias & Samiotou, forthcoming). Of course, as referred to in (Melamed, 2000): "bitext correspondence is typically only partial – many words in each text have no clear equivalent in the other text." *Figure 6* shows an example of word-alignment information which is originally displayed in different colours but due to the black and white printing we provide it with arrows.



Η Ιστιοπλοΐα στους Παραολυμπιακούς Αγώνες

Sailing in the Paralympics <noalign>

Figure 6. Example of Word-Alignment Information

2.4 Multilingual Linking

Multilingual linking is a unique feature of the ET which automatically connects, under defined conditions, entries through a common language entry. More specifically, it generates multilingual indirect links (i.e. links that are not imported as such through the user interface) for entries of a language that have no direct links (i.e. links that have been imported as such either manually or from a file or as alignment results) to other languages. This is possible due to the fact that the ET system is not pair-based but fully multilingual. For example, if (1) and (2) translation links are imported in the database then (3) translation link is automatically generated:

(1) Greek ⇔ English
Ολυμπιακοί Αγώνες ⇔ Olympic Games

(2) French ⇔ English
Jeux Olimpiques ⇔ Olympic Games

(3) Greek ⇔ French
Ολυμπιακοί Αγώνες<->Jeux Olympiques

2.5 Machine Translation

The units that have been left untranslated due to low alignment scores can be exported and machine translated by ET, with both MT and TM activated, using at least the already imported alignment results. If fuzzy matches are located then the system suggests its target language equivalent as the translation of the input unit. When no fuzzy match can be located for all or part of the input units, MT processing is activated to contribute in the translation of the remaining untranslated input unit. The MT results are automatically post edited by the TLV feature (see section 2.1) and imported, by filtering out the units that do not exist in the target pool of TM data.

3. The Olympic Games 2004 Project

The translation department of the Athens Organising Committee for the Olympic Games 2004 (ATHOC), selected ET to build TM data for Greek, English and French from the parallel texts they had previously translated, in order to generate as much feedback as possible form their legacy data. The legacy parallel texts were first processed on a monolingual level. Sentences and sub-sentences in English, French and Greek where imported in the TM together with their frequency of appearance information. A statistical repetition analysis on these texts indicated that the texts were quite repetitive, with a rate of 46% on both sentence and sub-sentence level.

Then, the legacy parallel texts were aligned. There were approximately 1,000 parallel document text pairs which resulted in approximately 15,000 TM sentences and 10,000 sub-sentences for each language pair. Source units were linked to one or multiple translations. More links where automatically generated through the multilingual linking feature of the ET.

ATHOC has been using ET in production since July 2003.

4. Conclusions

Parallel texts are a valuable resource for processing and extracting information for the translation process. ESTeam has proven that these resources are fully exploitable to improve any translation scenario where data is available. ESTeam has yet to explore the potential of parallel i.e. TM data as organisational and multilingual resource for knowledge representation.

References

Ahrenberg L., Andersson M. & Merkel M. (2000), A knowledge-lite approach to Word Alignment. In Veronis, J., Parallel Text Processing: Alignment and Use of Translation Corpora. (Kluwer Academic, 2000).

ESTeam AB. (2004). ESTeam Translator© White Paper. *URL: www.esteam.gr*

Gale, William A. and Kenneth W. Church. (1991). Identifying word correspondences in parallel texts. In Fourth DARPA Workshop on Speech and Natural Language, Asilomar, California.

Isabelle, Pierre, M. Dymetman, G. Foster, J-M. Jutras, E. Macklovitch, F. Perrault, X. Ren and M. Simard. (1993). Translation Analysis and Translation Automation. In Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation, Kyoto, Japan

Kranias L. (1995). A New Optimal Algorithm for the Solution of a Generalised Assignment Problem - Application in Automatic Text Alignment. Proceedings of the International Conference on Systems, Man & Cybernetics, Vancouver, Canada

Kranias Lambros, Samiotou Anna. (2004). Automatic Translation Memory Fuzzy Match Post-Editing: A Step beyond Traditional TM/MT Integration. Proceedings of LREC2004, Lisbon, Portugal.

Melamed, I. Dan. (1997). Automatic discovery of non-compositional compounds in parallel data. In Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP-97), Brown University.

Meyers A., Kosaka M. and Grishman R. (1998). A Multilingual Procedure for Dictionary-Based Sentence Alignment. Proceedings of ACL-COLING-98, Montreal, Canada.