Example-based Machine Translation via the Web

Nano Gough, Andy Way and Mary Hearne

School of Computer Applications,
Dublin City University,
Dublin, Ireland.
away@computing.dcu.ie

Abstract. One of the limitations of translation memory systems is that the smallest translation units currently accessible are aligned sentential pairs. We propose an example-based machine translation system which uses a 'phrasal lexicon' in addition to the aligned sentences in its database. These phrases are extracted from the Penn Treebank using the Marker Hypothesis as a constraint on segmentation. They are then translated by three on-line machine translation (MT) systems, and a number of linguistic resources are automatically constructed which are used in the translation of new input.
We perform two experiments on testsets of sentences and noun phrases to demonstrate the effectiveness of our system. In so doing, we obtain insights into the strengths and weaknesses of the selected on-line MT systems. Finally, like many example-based machine translation systems, our approach also suffers from the problem of 'boundary friction'. Where the quality of resulting translations is compromised as a result, we use a novel, *post hoc* validation procedure via the World Wide Web to correct imperfect translations prior to their being output to the user.

1 Introduction

Translation memory (TM) systems have rapidly become the most useful tool in the translator's armoury. The widespread availability of alignment software has enabled the creation of large-scale aligned bilingual corpora which can be used to translate new, unseen input. Many people believe that existing translations contain better solutions to a wider range of translation problems than other available resources (cf. Macklovitch, 2000). However, the main problem with these knowledge sources is that they are aligned only at sentential level, so that the potential of TM systems is being vastly underused.

This constraint on what segments can be aligned is overcome in Example-based Machine Translation (EBMT) systems. Like TM systems, EBMT requires an aligned bilingual corpus as a prerequisite, but translational correspondences can in addition be derived at sub-sentential level, which is not possible in TM systems. Accordingly, EBMT systems *generate* translations of new input by combining chunks from many translation examples; the best that TM software can do is to suggest the closest 'fuzzy' matches in its database for users to combine themselves in the formation of the translation.

In section 2, we show how the Marker Hypothesis (Green, 1979) can be used to create a 'phrasal lexicon' which renders the database of examples far more useful in building translations of new input. This lexicon was constructed by extracting over 200,000 phrases from the Penn Treebank and having them translated into French by three on-line machine translation (MT) systems. These three sets of translations were stored separately, and used as the basis for our EBMT system in translating two new testsets of NPs and sentences. As well as translating these examples using chunks from each of the individual sets of translations (A, B and C), in subsequent experiments we combined the three sets firstly in three new pairwise sets (AB, AC, BC), followed by combining them all together (ABC). The way that the chunks were combined and translations obtained is described in section 3. The results are presented in section 4.

Like many EBMT systems, our approach suffers from the problem of 'boundary friction'. Where the quality of resulting translations is compromised as a result, we use a novel, *post hoc* validation procedure via the World Wide Web, described in section 5, to correct imperfect translations prior to their being output to the user. Finally, in section 6 we conclude and outline some ideas for further research.

2 The Phrasal Lexicon

Other researchers have also noted this advantage of EBMT systems over TM software, namely the ability to avail of sub-sentential alignment. Simard and Langlais (2001) propose the exploitation of TMs at a sub-sentential level, while Schäler *et al.* (2002) describe a vision in which a phrasal lexicon occupies a central place in a hybrid integrated translation environment.

Our phrasal lexicon was built in two phases. Firstly, a set of 218,697 English noun phrases and verb phrases was selected from the Penn Treebank. We identified all rules occurring 1000 or more times and then eliminated those that were not relevant, e.g. rules dealing only with digits. Of the rules with a RHS containing a single non-terminal, only those rules whose LHS is VP were retained in order to ensure that intransitive verbs were represented in our database of translations. In total, just 59 rules out of a total of over 29,000 were used in creating the lexicon.

These extracted English phrases were then translated using three different on-line MT systems:

- SDL International's Enterprise Translation Server[1] (system A)
- Reverso by Softissimo[2] (B)
- Logomedia[3] (C)

These MT systems were selected as they enable batch translation of large quantities of text. We found that the most efficient way to translate large

[1] `http://www.freetranslation.com`
[2] `http://trans.voila.fr`
[3] `http://www.logomedia.net`

amounts of data via on-line MT systems was to send each document as an HTML page where the phrases to be translated are encoded as an ordered list. The English phrases were therefore automatically tagged with HTML codes and passed to each translation system via the Unix 'wget' function. This function takes a URL as input and writes the corresponding HTML document to a file. If the URL takes the form of a query then the document retrieved is the result of the query, namely the translated web page. Once this is obtained, retrieving the French translations and associating them with their English source equivalents is trivial.

Despite the (often) poor output obtained from these systems, impressive results may still be obtained. We do not validate the translations prior to inserting them into our databases. Of course, if we were to do so, or use 'better' systems, then the results presented in section 4 would improve accordingly.

2.1 The Marker Lexicon

In their *Gaijin* system, Veale and Way (1997) propose the use of the Marker Hypothesis to create aligned chunks at sub-sentential level. The Marker Hypothesis is a psycholinguistic constraint on grammatical structure that is minimal and easy to apply. Given that it is also arguably universal, it is clear to see that it has obvious benefits in the area of translation.

The Marker Hypothesis states that all natural languages contain a closed set of specific lexemes and morphemes which indicate the grammatical structure of strings. As in *Gaijin*, we exploit such lists of known marker words for each language to indicate the start and end of segments. For English, our source language, we use the six sets of marker words in (1), with a similar set produced for French, our target language:

(1)
Det: {the, a, an, those, these, ...} Conj: {and, or, ...}
Prep: {in, on, out, with, to, ...} Poss: {my, your, our, ...}
Quant: {all, some, few, many, ...} Pron: {I, you, he, she, it, ...}

In a pre-processing stage, the aligned sentence pairs are traversed word by word, and whenever any such marker word is encountered, a new chunk is begun, with the first word labelled with its marker category (Det, Prep etc.). The following example illustrates the results of running the marker hypothesis over the phrase *on virtually all uses of asbestos*:

(2) `<PREP>` on virtually, `<QUANT>` all uses, `<PREP>` of asbestos

In addition, each chunk must also contain at least one non-marker word, so that the phrase *out in the cold* will be viewed as one segment, rather than split into still smaller chunks.

For each $\langle English, French_X \rangle$ pair, where X is one of the sets of translations derived from separate MT systems (A, B, and C), we derive separate marker lexicons for each of the 218,697 source phrases and target translations. Given that English and French have essentially the same word order, these marker lexicons are predicated on the naïve yet effective assumption that marker-headed chunks in the source S map sequentially to their target equivalents T, i.e. $\text{chunk}_{S1} \longrightarrow \text{chunk}_{T1}$, $\text{chunk}_{S2} \longrightarrow \text{chunk}_{T2}, \ldots \text{chunk}_{Sn} \longrightarrow \text{chunk}_{Tn}$. Using the previous example of *on virtually all uses of asbestos*, this gives us:

(3)
<PREP> on virtually : sur virtuellement
<QUANT> all uses : tous usages
<PREP> of asbestos : d'asbeste

In addition, we generalize over the phrasal marker lexicon along the lines of (Block, 2000). Taking (3) as input, we produce the templates in (4):

(4)
<PREP> virtually : <PREP> virtuellement
<QUANT> uses : <QUANT> usages
<PREP> asbestos : <PREP> asbeste

This allows other marker words of the same category to be substituted for those in the phrasal chunks. For instance, in our testset of NPs, we do not locate *the fully operational prototype*, the nearest approximation being *a fully operational prototype*. By replacing the marker word *a* with <DET>, we can search the generalized lexicon for the chunk <DET> *fully operational prototype*, retrieve its translation and insert translations for *the*. Errors of agreement in this insertion process may again be corrected using the techniques involved in section 5.

Finally, we take advantage of the further assumption that where a chunk contains just one marker word in both source and target, these words are translations of each other. Where a marker-headed pair contains just two words, therefore, we are able to extract a further bilingual dictionary. From the chunks in (3), we can extract the following six word-level alignments:

(5)
<PREP> on : sur <LEX> virtually : virtuellement
<QUANT> all : tous <LEX> uses : usages
<PREP> of : d' <LEX> asbestos : asbeste

That is, using the marker hypothesis method, smaller aligned segments can be extracted from the phrasal lexicon without recourse to any detailed parsing techniques. When matching the input to the corpus, we search for chunks in the order (original) phrasal dictionary $\longrightarrow$ phrasal marker lexicon (cf. (3)) $\longrightarrow$ generalized phrasal marker lexicon (cf. (4)) $\longrightarrow$ word-level marker lexicon (cf. (5)), so that greater importance is attributed to longer chunks, as is usual in most EBMT systems. The word for word translation pairs are only used when a translation cannot be formed in any other way.

Given that verbs are not a closed class, we take advantage of the fact that the initial phrasal chunks correspond to rule RHSs. That is, for a rule in the Penn Treebank VP $\longrightarrow$ VBG, NP, PP, we are certain (if the taggers have done

their job correctly) that the first word in each of the strings corresponding to this RHS is a VBG, i.e. a present participle. In such cases we also tag such words with the <LEX> tag, e.g. '<LEX> expanding : augmente'.

3 Chunk Retrieval and Translation Formation

In section 4, we describe two experiments, one on NPs and one on sentences. In this section, we describe the processes involved in retrieving appropriate chunks and forming translations for NPs only, these being easily extensible to sentences.

3.1 Segmenting the Input

In order to optimize the search process, a given NP is segmented into smaller chunks. The system then attempts to locate these chunks individually and to retrieve their relevant translation(s). We use an n-gram based segmentation method, in that all possible bigrams, trigrams and so on are located within the input string and subsequently searched for within the relevant knowledge sources. [4]

3.2 Retrieving Translation Chunks

We use translations retrieved from three different sources A, B and C. These translations are further broken down using the Marker Hypothesis, thus providing us with an additional three knowledge sources A′, B′ and C′—the phrasal marker lexicons. These knowledge sources can be combined in several different ways. We have produced translations using information from a single source (i.e. A/A′, B/B′ and C/C′), pairs of sources (i.e. A/A′ & B/B′ (=AB), A/A′ & C/C′ (=AC), and B/B′ & C/C′ (=BC)), and all available knowledge sources (i.e. A/A′ & B/B′ & C/C′ (=ABC)). Each time a source language (SL) chunk is submitted for translation the appropriate target language (TL) chunks are retrieved and returned with a weight attached.

3.3 Calculation of Weights

We use a maximum of six knowledge sources: firstly, three sets of translations (A, B and C) retrieved using each on-line MT system; and secondly, three sets of translations (A′, B′ and C′) acquired by breaking down the translations retrieved at the initial stage using the Marker Hypothesis. Within each knowledge source, each translation is weighted according to the following formula:

[4] Of course, given our segmentation method, many of these n-grams cannot be found, given that new chunks are placed in the marker lexicon when a marker word is found in a sentence. Taking the NP *the total at risk a year* as an example, chunks such as 'the total at risk a' or 'at risk a' cannot be located, as new chunks would be formed at each marker word, so the best that could be expected here might be to find the chunks <DET> *the total*, <PREP> *at risk*, <DET> *a year* and recombine their respective translations to form the target string. In ongoing work, we are continuing to eliminate all such n-grams which are impossible to find from the search process.

(6) $$\text{weight} = \frac{\textit{no. occurrences of the proposed translation}}{\textit{total no. translations produced for SL phrase}}$$

For the SL phrase *the house*, assuming that *la maison* is found 8 times and *le domicile* is found twice, then P(*la maison* | *the house*) = 8/10 and P(*le domicile* | *the house*) = 2/10. Note that since each SL phrase will only have one proposed translation within each of the knowledge sources acquired at the initial stage, these translations will always have a weight of 1.

If we wish to consider only those translations produced using a single MT system (e.g. A and A′), then we add the weights of translations found in both knowledge sources and divide the weights of all proposed translations by 2. For the SL phrase *the house*, assuming P(*la maison* | *the house*) = 5/10 in knowledge source A and P(*la maison* | *the house*) = 8/10 in A′, then P(*la maison* | *the house*) = 13/20 over both knowledge sources. Similarly, if we wish to consider translations produced by all three MT systems, then we add the weights of common translations and divide the weights of all proposed translations by 6.

When translations have been retrieved for each chunk of the input string, these translated phrases must then be combined to produce an output string. In order to calculate a ranking for each TL sentence produced, we multiply the weights of each chunk used in its construction, thus favouring translations formed via larger chunks. Where different derivations result in the same TL string, their weights are summed and the duplicate strings are removed.

4 Experiments and System Evaluation

4.1 Experiment 1: Sentences

The automatically generated testset comprised 100 sentences, with an average length of 8.5 words (min. 3 words, max. 18).[5] The sentences were segmented using the *n*-gram approach outlined in section 3. Following the submission of these sentences to each of the knowledge sources, translations were produced for 92% of cases for systems A and C, and 90% for system B. The same 8 sentences fail to be translated by any of the systems (or combinations of systems) owing to a failure to locate a word within the word-level lexicon. For 48% of the successful cases, the translation was produced by combining chunks found in either the original phrasal lexicon or the phrasal marker lexicon. In 28% of cases, the translation was produced by locating single words in the word-level lexicon and inserting these into the translation at the correct position. The remaining 16% of translations were produced with recourse to the generalized marker lexicon.

[5] The testset itself adversely affected the results derived from this experiment. Given the preference for on-line systems to process S-level expressions, third person plural dummy subjects were provided. As a consequence, the VPs in our phrasal lexicon are for the most part in this corresponding form also. The majority of subject NPs in our sentence testset are singular, which almost guaranteed a lower quality translation. Nevertheless, the results achieved are still reasonable and can easily be improved by adding new, relevant translation examples to the system database.

While coverage is important, the quality of translations produced is arguably more important. All translations produced were evaluated by two native speakers of French with respect to the following classification schema:

- Score 3: contains no syntactic errors and is intelligible;
- Score 2: contains (minor) syntactic errors and is intelligible;
- Score 1: contains major syntactic errors and is unintelligible.

The results obtained are shown in Table 1. For the majority of those translations assigned a score of 2, the verb was either in the incorrect form, or the agreement between noun and verb was incorrect. Most of these examples may be corrected using the *post hoc* validation procedure outlined in section 5.

System	Score 1	Score 2	Score 3
A	14.2%	51.2%	34.6%
B	8.9%	54.7%	36.4%
C	4.4%	59.1%	36.5%

Table 1. Quality of Translations obtained for Sentence Testset

Like many other data-driven approaches to translation, our EBMT system produces many translation candidates for the user's perusal. Another important issue, therefore, is that the 'correct' translation be as near to the top among those translations in which the system has the most confidence (i.e. the 'best' translation). We discuss issues pertaining to combining chunks from different on-line systems in section 4.3. For the individual systems, however, in over 65% of cases the 'correct' translation was ranked first by the system, and in all cases the 'correct' translation was located in the top five-ranked translations.

4.2 Experiment 2: Noun Phrases

A second experiment employing a testset of 200 noun phrases was subsequently undertaken. Here the average NP length was 5.37 words (min. 3 words, max. 10). In 94% of cases (188 NPs), at least one translation was produced for either system A, B or C. On average, about 54% of translations are formed by combining chunks from the phrasal lexicon, about 9% are produced by searching the generalized chunks, and about 37% are generated by inserting single words from the word-level lexicon at the appropriate locations in phrasal chunks. The failure to produce a translation in 6% of cases was invariably due to the absence of a relevant template in the generalized marker lexicon.

The translated NPs were again evaluated using the scale outlined in the previous section. The results achieved are summarized in Table 2, and are somewhat more definitive than with the sentence testset summarized in Table 1. Our EBMT system works best with chunks derived from system C, Logomedia. with a clear 7% more translations with no errors, and only 2.6% of translations deemed unintelligible. System B again outperforms System A.

System	Score 1	Score 2	Score 3
A	11.9%	51.4%	36.7%
B	4.8%	53.8%	41.4%
C	2.6%	49%	48.4%

Table 2. Quality of Translations obtained for NP Testset

As was the case with sentences, our EBMT system produces many translation candidates for NPs. For instance, the NP *a plan for reducing debt over 20 years* receives 14 translations using chunks from system A, 10 via B and 5 via C. When we combine chunks from more than one system, this rises to 224 for ABC. For the individual systems, in almost all cases, the 'correct' translation was located within the top five ranked translations proposed by the system, and at worst in the top ten.

4.3 Extending the Experiments

We also examined the performance of our EBMT system on both testsets when it has access to chunks from more than one system. We performed four more experiments—three pairwise comparisons (AB, AC, BC) and one threefold (ABC).

With respect to coverage, all four combinations translated 92% of the sentence testset. For the NP testset, in the pairwise comparison, coverage ranged from 94% (AB) to 95.5% (both AC and BC), while ABC translated 96% of the NPs.

We also evaluated translation quality using the same 3-point scale. For sentences, we observed that chunks involving some combination of system C perform better (AC and BC both achieving 48.9% top score, compared with AB's 47.2%), with ABC outperforming any of the pairwise systems (50% of translations scoring 3). On the NP testset, AC (62.8% top score) and BC (62.3%) both outperform AB (58%), while ABC scores 3 for 70.8% of NPs, with only 0.5% (i.e. one NP) regarded as unintelligible.

Regarding the relative location of the 'correct' translation for sentences, the 'correct' translation is to be found in the top ten-ranked translations in all permutations of combinations of chunks, with at least 97.3% found in the top five and 54% ranked first. For NPs, the 'correct' translation is to be found in the top five-ranked translation candidates in almost all cases.

5 Validation and Correction of Translations via the Web

A translation can only be formed in our system when the recombination of chunks causes the input NP to be matched exactly. Therefore, if all chunks are not retrieved then no translation is produced. When a translation cannot be produced by combining the existing chunks, the next phase is to check whether a translation can be formed by the insertion of single marker words into the target string. Given the NP *the personal computers*, this can be segmented into three

possible chunks: *the personal, personal computers* and *the personal computers*. The chunk *personal computers* is the only one retrieved in the phrasal lexicon of our system. As it does not match the input NP exactly, its translation does not qualify as a complete translation, of course. The system stores a list of marker words and their translations in the word-level marker lexicon. A weight derived from the method in (6) is attached to each translation. The system searches for marker words within the string and retrieves their translations. In this case, the marker word in the string is *the* and its translation can be one of *le, la, l'* or *les* depending on the context. The system simply attaches the translation with the highest weight to the existing chunk (*ordinateurs personnels*) to produce the translation *la ordinateurs personnels*. Of course, the problem of boundary friction is clearly visible here.

However, rather than output this wrong translation directly, we use a *post hoc* validation and (if required) correction process based on (Grefenstette 1999). Grefenstette shows that the Web can be used as a filter on translation quality simply by searching for competing translation candidates, and selecting the one which is found most often. Rather than search for competing candidates, we select the 'best' translation and have its morphological variants searched for on-line. In the example above, namely *the personal computers*, we search for *les ordinateurs personnels* versus the wrong alternatives *le/la/l'ordinateurs personnels*. Interestingly, using Altavista, and setting the search language to French, the correct form *les ordinateurs personnels* is uniquely preferred over the other alternatives, as it is found 980 times while the others are not found at all. In this case, this translation overrides the 'best' translation *la ordinateurs personnels* and is output as the final translation. This process shows that while the Web is large, despite the fact that it is unrepresentative and may be seen to contain what might be considered 'poor quality' data, it remains a resource which is of great use in evaluating translation candidates.

6 Conclusions and Further Work

We have presented an EBMT system based on the Marker Hypothesis which uses *post hoc* validation and correction via the Web. A set of over 218,000 NPs and VPs were extracted automatically from the Penn Treebank using just 59 of its 29,000 rules. These phrases were then translated automatically by three on-line MT systems. These translations gave rise to a number of automatically constructed linguistic resources: (i) the original ⟨*source,target*⟩ phrasal translation pairs; (ii) the phrasal marker lexicon; (iii) the generalized phrasal marker lexicon; and (iv) the word-level marker lexicon. When confronted with new input, these knowledge sources are searched in turn for matching chunks, and the target language chunks are combined to create translation candidates.

We presented two experiments which showed how the system fared when confronted with NPs and sentences. For the former, we translated 96% of the testset, with 71% of the 200 NPs being translated correctly, and 99.5% regarded as acceptable. For our 100 sentences, we obtained translations in 92% of cases,

with a completely correct translation obtained 50% of the time, and an acceptable translation in 96.8% of cases. Importantly, the 'correct' translation was to be found in almost all cases in the top five-ranked translation candidates output by our system. Prior to outputting the best-ranked translation candidate, its morphological variants are searched for via the Web in order to confirm it as the final output translation or to propose a corrected alternative.

A number of issues for further work present themselves. The decision to take all rules occurring 1000 or more times was completely arbitrary and it may be useful to include some of the less frequently occurring structures in our database. Similarly, it may be a good idea to extend our lexicon by including more entries using Penn-II rules where the RHS contains a single non-terminal.

Furthermore, the quality of the output was not taken into consideration when selecting the on-line MT systems from which all our system resources are derived, so that any results obtained may be further improved by selecting a 'better' MT system which permits batch processing.

Finally, we want to continue to improve the evaluation of our system, firstly by experimenting with larger datasets, and also by removing any notion of subjectivity by using automatic evaluation techniques.

In sum, we have demonstrated that using a 'linguistics-lite' approach based on the Marker Hypothesis, with a large number of phrases extracted automatically from a very small number of the rules in the Penn Treebank, many new reusable linguistic resources can be derived automatically which can be utilised in an EBMT system capable of translating new input with quite reasonable rates of success. We have also shown that the Web can be used to validate and correct candidate translations prior to their being output.

References

1. Block, H. U.: Example-Based Incremental Synchronous Interpretation. In Wahlster, W. (ed.) *Verbmobil: Foundations of Speech-to-Speech Translation*, Springer-Verlag, Berlin Heidelberg New York (2000) 411–417
2. Green, T.R.G.: The Necessity of Syntax Markers. Two experiments with artificial languages. *Journal of Verbal Learning and Behavior* **18** (1979) 481–496
3. Grefenstette, G.: The World Wide Web as a Resource for Example-Based Machine Translation tasks. In *Proceedings of the ASLIB Conference on Translating and the Computer* **21**, London (1999)
4. Macklovitch, E.: Two Types of Translation Memory. In *Proceedings of the ASLIB Conference on Translating and the Computer* **22**, London (2000)
5. Schäler, R., Carl, M., Way, A.: Example-Based Machine Translation in a Hybrid Integrated Environment. In Carl, M., Way, A. (eds.) *Recent Advances in Example-Based Machine Translation*, Kluwer Academic Publishers, Dordrecht, The Netherlands (in press) (2002)
6. Veale, T., Way, A.: Gaijin; A Bootstrapping, Template-Driven Approach to Example-Based MT. In *Proceedings of the Second International Conference on Recent Advances in Natural Language Processing*, Tzigov Chark, Bulgaria (1997) 239–244