Example-Based Machine Translation using Connectionist Matching

Ian J. McLean

Centre for Computational Linguistics
UMIST, PO Box 88
Manchester M60 1QD, England.
ian@ccl.umist.ac.uk

Abstract

This paper proposes an alternative approach to matching input text with example text in an Example-Based Machine Translation system. The approach employs a connectionist network to compute a measure of distance between the input text and the source members of source / target text pairs contained in a bilingual corpus.

Keywords: Empiricist; Connectionism

Introduction

A framework for Example-Based Machine Translation (EBMT) was first proposed by Nagao [9] where he suggests that rather than requiring deep linguistic analysis, translation is achieved by

> "... properly decomposing an input sentence into certain fragmental phrases [...], then by translating these fragmental translations into other language phrases and finally by properly composing these fragmental phrases into one long sentence." (p. 179)

This paper will address a problem inherent in the example-based approach to MT: the selection from a set of examples (a bilingual corpus) the most suitable translation pair(s) given an input in the source language, a process commonly referred to as *matching.* This selection is often performed by first computing a *distance measurement* between the input to be translated and the source language examples. Figure 1 illustrates the architecture within which this process operates. The matching process selects from a set of translation pairs the pairs whose source sentence is the

smallest distance from an input sentence. Once suitable translation pairs have been selected

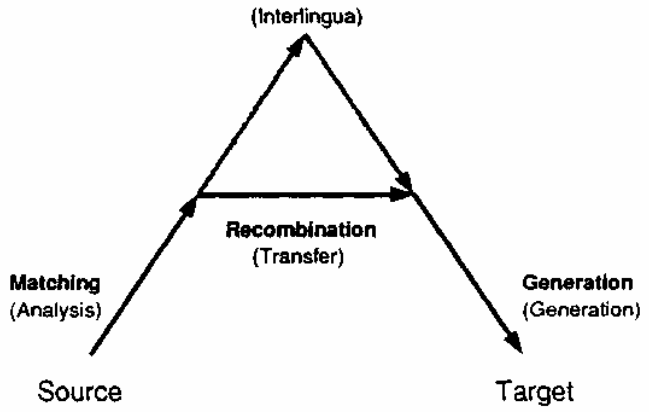

Figure 1 **EBMT Architecture.** The figure shows EBMT architecture with conventional terminology shown in parentheses.

by the matching process, constituents from the matched examples are recombined in order to more closely match the input sentence. Finally, a translation is generated using the translations of the recombined source sentence constituents. For further explanation see Hutchins & Somers [5] (pp. 125-130).

A number of approaches to distance measurement have been proposed which employ both statistical and heuristic techniques. Sato and Nagao [12] combine two heuristics in a similarity measure which accounts for both the size of a translation unit and the 'environmental similarity'[1] of the input and source example. The latter is computed with the aid of a thesaurus which specifies *similarity values* between words in the same language e.g. the similarity between *book* and *notebook* is defined to be 0.8.

Similarly, Sumita *et al.* [14] employ a thesaurus containing an abstraction hierarchy to which semantic distances (relative to the root) are attached at the nodes. Attributes from the input are matched to attributes in the thesaurus and for each pair a common level of semantic abstraction is identified. The semantic distance from that level is taken as a measure of semantic distance between the two attributes. A weight derived from this distance is combined with a frequency based *syntactic* distance measure and the sum of all of the attribute level products provides the final distance measurement for the whole translation.

The 'Bilingual Knowledge Bank' proposed in [11] following the DLT project also uses a bilingual thesaurus, two interconnected Textual Knowledge Banks' (TKBs), which contains aligned 'translation units' generated by a conventional parsing process. Matching in this approach originally involved the rule-based parsing of an input text into a dependency tree although this process was subsequently developed to become purely analogical.

The approaches taken by Sumita *et al.,* Sato and Nagao and Sadler rely upon the use of extensive thesauri, requiring that formal linguistic information be encoded in a 'rationalist' manner and thus retaining the problem of rigidity implicit in such static definitions. It may also be argued (aside from any practical considerations) that the 'formal' embodiment of this type of information in a thesaurus runs contrary to the ethos of the example-based paradigm which they employ

[1] Environmental similarity is a measure of the syntactic context in which a translation unit may occur.

elsewhere in their work: the decomposition of input sentences into word dependency trees in [12], the analysis phase mentioned in [14] and the parsing process required for the initial stages of TKB construction in [11].

Carroll [1] introduces the concept of an *angle of similarity as* a measure of distance between sentences. This angle is calculated using a triangle whose three points represent the two sentences being compared and a 'null sentence'. The length of the sides from this null point to the points representing the two sentences are the respective sizes of those sentences and the length of the third side is the difference between the two. The size of a sentence is calculated by costing the add, delete and replace operations necessary to derive one sentence from the other using costs from a set of 'rules' embodied in the system. Carroll shows that the angle at the null sentence point (between the adjacent and hypotenuse sides) provides a measure of distance which reduces 'undesirable' length effects whereby a sentence may be selected from a set of examples by virtue of the proximity of its length to the input sentence length rather than its qualitative or directional similarity to other, longer, examples. However, the derivation process requires the formal specification of *ad hoc* cost measurement rules which define the cost of the basic operations it employs.

Jones [6] proposes an approach employing an *analogical modelling* technique whereby the distance between an input and example is calculated by a comparison of feature vectors attached to the input and examples. A number of these vectors may be attached to each example, describing features at different levels: morphological and clausal, for example. From the example database, examples are selected for both the similarity of their feature vectors with the input and for the similarity between their 'outcomes'[2] and the outcome predicted by the closest analogy to the input. The probability that an example will serve as the analogical model is calculated based upon its similarity and frequency of occurrence in the example database (for a complete explanation of analogical modelling see Skousen [13]). While this approach does not require an extensive thesaurus, it is necessary for a corpus to be augmented with the feature information required by the analogical modelling process in order to produce the example database. However, the extraction of this linguistic information may be readily automated and is not as extensive a problem as either the encoding of linguistic rules or generation of *ad hoc* thesaurus entries.

In the following sections a connectionist[3] alternative to the above which avoids the use of extensive thesauri and alleviates the extent of corpus augmentation will be outlined. It will be shown experimentally that connectionism offers an alternative to the above solutions to distance measurement in that it may account for length, frequency, syntactic and semantic contextual effects and it will be concluded that on the basis these results the application of connectionism to MT should be pursued further.

Architecture

The proposed approach employs a two layer connectionist architecture (figure 2). As its various names suggest, the connectionist paradigm is based upon the use of a large number of very simple interconnected processing *units* (these are sometimes, rather controversially, referred to as neurons). A pattern of *activation* (often binary) is presented to the source units and this activation is propagated through weighted connections (the weights of which are initially random) to the units at the end of the connections which combine the weighted activations from their incoming connections by way of a *combination function* (usually summation) to produce a *net* value for that unit. The net value is processed by an *activation function* associated with the unit to produce its new activation. The key to the operation of such a network is the relative strengths of

[2] For example at the morphological level the outcome of *the* would be Det and at the clausal level the outcome of *the dog* would be NP.

[3] The *connectionist* paradigm is also referred to as Parallel Distributed Processing (PDP) or the use of neural networks.

its interconnecting weights. These are modified during an initial learning phase where the network is provided with not only a source pattern but the target pattern which it is expected to make in response[4]. By comparing the actual and desired target responses, a *learning algorithm* adjusts the weights of the connections in proportion to a learning rate and, given a suitably stable training set, the weights stabilize. Once the weights have become stable the network has learned its training set and with the learning 'turned off (i.e. learning rate = 0) can be presented with patterns which it will attempt to classify but not learn.

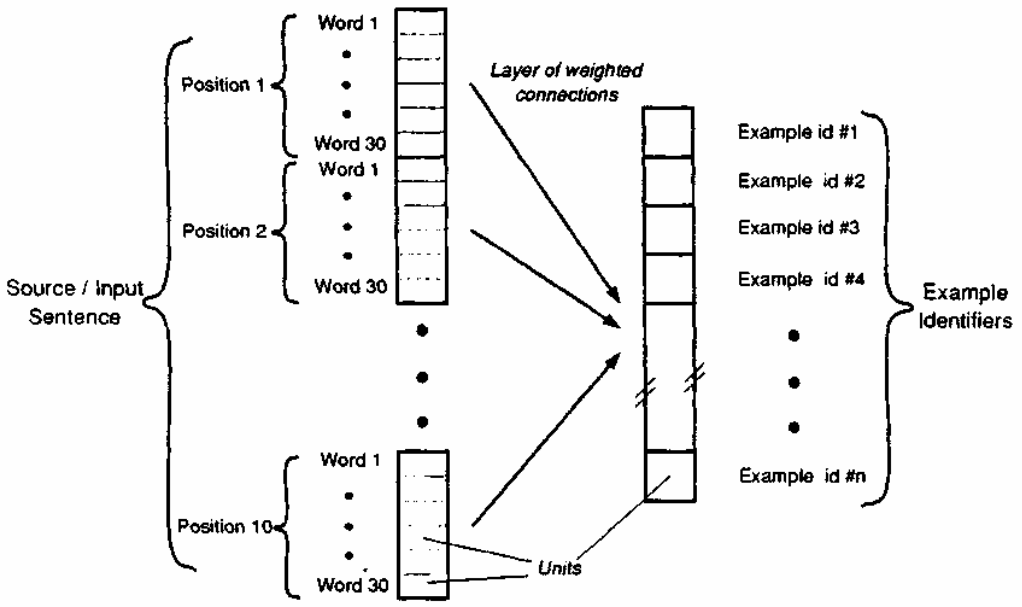

Figure 2 Network Architecture.

In the proposed architecture, units on the input are grouped so that for each word position in the input sentence there are 30 units, each unit corresponding to a word. Consequently only one unit in each group may be active at any given time as only a single word can occupy a position in a sentence at one time. Output units are encoded differently with a single unit corresponding to a translation. Thus the network is configured to map from source sentences with word-level resolution to targets with sentence-level resolution which enables it to learn to select a translation by virtue of detailed source-level information. Although the detail employed in this particular model is at the word level, the architecture allows for the use of encoded morphological, clausal and semantic information (see discussion in the summary). The network weights fully interconnect the input units to the output units in a unidirectional manner and are adjusted using the Delta learning algorithm described by Rumelhart & McClelland [10] (chapter 11).

Experiments

In each of the following experiments the same network and training set was employed. The training set comprised 32 phrases taken from business letters and was of the form

{<source sentence , <unique example identifier>}

[4] **Some networks are unsupervised and do not require a target pattern to learn but this type of network is beyond the scope of this paper.**

e.g.

{[the, cat, sat, on, the, mat], 1).

The unique example identifier can be considered to be a pointer to the corresponding translation pair in the example set of the form

<unique example identifier> —> {e_s, e_t)

e.g.

1 —> {[the, cat, sat, on, the, mat], [le, chat, assis, sur, le, tapis]}

The learning rate used in the Delta learning algorithm was 0.5 and the activations contained on the input and target vectors were binary.

The network was trained by presenting source sentence and unique translation pair identifier patterns to the input and output units of the network respectively and applying the learning algorithm to adjust the weights of the connections. This was iterated over each of the 32 entries in the training set and the network was exposed to 65 sets of iterations before learning was complete and the tests performed.

1 Salient Feature Identification

The first of the experiments demonstrates the network's ability to identify salient features of the input sentences. Using the sentences (1) and (2) the ability of the network to identify salient features (words) from an input pattern (sentence) can be demonstrated.

(1) We cannot accept your conditions of payment.
(2) We cannot accept your conditions of delivery.

After training, the weights connecting units representing words in position 7 in the source sentence pattern[5] to units representing the unique example identifier are as shown in figure 3. It can be seen that there is a strong positive weight from the *payment* unit and a strong negative weight from the *delivery* unit into the target unit representing the example (1). In a similar manner, the converse is true for the weights into the target unit representing the example (2).

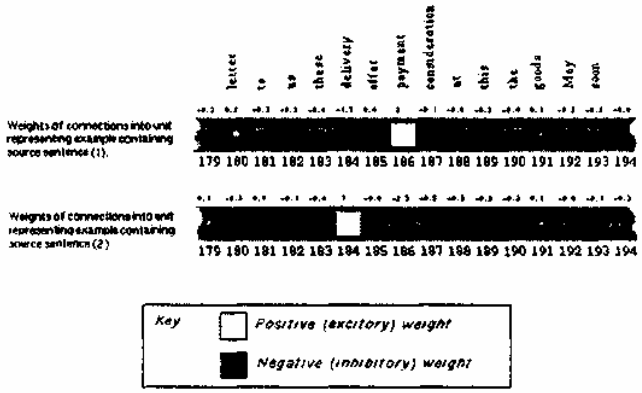

Figure 3 **Weights indicating salient features.** The figure shows the weights of the connections from units 180 through 194 (which encode words in the source sentences occupying the 7th position in those sentences) to the units encoding the translation pairs (1) and (2). The size of the weight is proportional to the size of the square representing it and its sign indicated by the square's colour.

[5] The words which may appear in position 7 of the sentence are encoded on units 180 to 209 of the network.

2 Graceful Degradation — 'Best Guess'

An inherent property of connectionist networks is their ability to make educated guesses and is exploited in models which need to generalize about pattern classifications or withstand noisy data or even physical damage. In this particular test the network, having been taught the unique example identifier for the sentence

(3) We acknowledge receipt of your letter.

is able to make a generalization to the effect that any pattern *similar* to (3) should invoke the same response as (3). So, for example, when shown (but not *taught)* the sentence

(4) We acknowledge receipt of your memo.

the response of the target unit representing the example containing sentence (3) is reduced from 0.91 to 0.82 reflecting the slight dissimilarity between the two. However, where the network has been taught two similar sentences, for example

(5) The delivery time is 4 months.

(6) The delivery time is 8 months.

and subsequently shown a third

(7) The delivery time is 7 months.

the response of the target unit representing the example containing sentence (5), for example, drops from 0.84 to 0.36. This drop is more significant than that exhibited when sentence (4) was shown to the network as sentence (7) differs from *two* sentences by only one word. Thus the network is 'hedging its bets' by activating *both* of the target units corresponding to the examples containing sentences (5) and (6) respectively.

3 Length Effects

It can also be shown that the network is able to reflect differences in sentence length in the responses it makes. Table 1 shows the responses produced by target units corresponding to the examples containing sentences (8) and (9) when subsequently shown sentences (8) and (9) which were contained in the training set and sentence (10) which was not. The responses to sentence (10) illustrate two points: first the network is attempting to select the most likely translation pair for a sentence which it has never seen before and is consequently making an informed guess (see previous experiment); and second there is a stronger response from the target unit corresponding to the example containing sentence (8). This is because the lengths of sentences (8) and (10) differ by only one word whereas those for sentences (9) and (10) differ by two.

Sentence	Response (8)	Response (9)
(8) We can deliver the goods.	0.80	0.05
(9) We can deliver the goods earlier than agreed.	0.05	0.93
(10) We can deliver the goods earlier.	0.52	0.39

Table 1 **Responses reflecting length effects.**

4 Frequency Effects

In this experiment, the original corpus is augmented with two extra occurrences of the sentence

(11) we do not need these items at present.

and the network re-trained. The post-learning responses to this sentence is shown in table 2 together with those for the sentence

(12) we do not stock these items at present.

which occurred only once in the corpus and the sentence

(13) we do not have these items at present.

which did not occur in the corpus at all. The table shows that although the previously unseen

Sentence	Response (11)	Response (12)
(11)	0.9	0.09
(12)	0.2	0.77
(13)	0.52	0.37

Table 2 **Frequency effects upon example selection.**

sentence (13) differs from sentences (14) and (15) by a single word, the response representing the example containing source sentence (14) is more highly active because of sentence (16)'s higher frequency of occurrence. Although the network shows a bias in favour of the response unit representing example (11), it does *not* reflect the 3:1 relationship implied by the frequency with which the examples are presented to the network during training (see discussion in the summary below).

Summary of Results

The results above provide a simple illustration of the suitability of connectionist networks for the matching process required in EBMT. They show that a network may be taught a set of example translations and from these select the most appropriate translation for a previously unseen source sentence. The model used in the experiments above accounts for a number of the factors employed in existing measurements.

In common with the approaches taken by Sato & Nagao [12], Sumita *et* a/. [14] and Jones [6], the connectionist model reflects a degree of confidence in the translation which it has selected (the reciprocal of which is referred to as a distance). As noted by Skousen [13] (p.81) connectionist networks do not produce probabilities which directly reflect the frequency at which a given response is produced because of 'interference' between weight changes made for different source/target pairs[6]. However, as a means of establishing the suitability of one translation over another, connectionism remains a useful paradigm to employ.

Despite the probability problem outlined above, it was shown in experiment 4 that the connectionist model also exhibits frequency effects. The bias illustrated by the data in table 2

[6] **This is a fundamental feature of connectionism which gives rise to emergent properties such as noise tolerance and multiple constraint satisfaction.**

is similar in nature to the frequency effect of the analogical modelling approach in that as the frequency of occurrence of a feature in the training set increases, so does the likelihood that its corresponding translation will be selected.

The connectionist approach is also sensitive to sentence length in a similar way to that employed in Sato & Nagao's [12] distance measurement. However, it is clear that using the architecture proposed above this length sensitivity relies upon the absolute positioning of words in the sentence. Indeed, the correct operation of the whole network is dependent upon this absolute positioning (see Conclusion section).
The length effects shown by the model in its current form differ from those outlined in [1] in that equal favour is given to sentence length and qualitative similarity. This may be achieved in a connectionist model by the introduction of semantic data (see below) to provide a qualitative bias or by changes in the combination function to provide an explicit bias in favour of words occurring at the beginning of a sentence. Thus a previously unseen short sentence would carry almost as much weight as a similar long one of which it was a fragment.

The architecture employed in the above experiments is structured to process data at word and clausal levels. In reality, however, the network is simply learning to map one pattern to another and consequently these patterns may be augmented with encoded syntactic features or *semantic microfeatures* like those used by McClelland & Kawamoto [7] (Ch.19) to enable the network to select translations based upon syntactic and semantic cues in a similar way to Sato & Nagao's [12] *syntactic context* and the *semantic distance* of Sumita *et al.* [14].

Conclusion

A simple connectionist architecture for distance measurement has been proposed and the results presented above illustrate that the approach is capable of reproducing some of the desirable features of existing measurements. Connectionism is inherently empirical and its application to EBMT seems inevitable. Furthermore, the approach not only represents a shift away from inflexible rule-based rationalist techniques but also from conventional iterative computation with all the subsequent advantages that parallel computation has to offer[7].

However, there exists one major problem with the model described in this paper, which is its total reliance upon the absolute positioning of words in a sentence. For example it would fail to recognize any similarity between the following sentences (given that, say, sentence (19) was included in the training set and (20) was not):

(19) John walked through the door.
(20) Mr. Smith walked through the door.

The system's reliance upon absolute positioning means that having been trained to recognize (19) and subsequently shown (20) it will attempt to match *John* with *Mr., walked* with *Smith* and so on. Thus the fact that the syntax of the two sentences is virtually identical would not be reflected by the system's response to the previously unseen sentence (20). A solution to this problem lies in the move from the *spatial* domain to the *temporal* domain. It is clear that language processing in humans whether it be spoken or written, comprehension or translation, is in the latter of the two domains as speech is obviously time based as is the scanning of the words on a written page. Having demonstrated that the application of the connectionist paradigm to the problem of EBMT shows potential it is now possible to look to more complex models [3][2][4][8] to address the outstanding problems faced by connectionist machine translation.

[7] An understanding of parallel computation is not required to appreciate this point. The only model we have of a fully operational translation system (the human brain) employs massive parallelism.

Bibliography

[1] Jeremy J. Carroll. *Repetitions processing using a Metric Space and the Angle of Similarity.* Technical Report No. 90/3. Centre for Computational Linguistics, UMIST, Manchester, 1990.

[2] A. Cleeremans, D. Servan Schreiber, and J.L. McClelland. Finite state automata and simple recurrent networks. *Neural Computation,* 1:372-381, 1989.

[3] J.L. Elman. Finding structure in time. *Cognitive Science,* 14:179-211, 1990.

[4] G. Houghton. The problem of serial order: A neural network of sequence learning and recall. In R. Dale, C. Melish, and N. Zock, editors, *Current Research in Natural Language Generation,* pages 287-320. London Academic Press, 1990.

[5] W.J. Hutchins and H.L Somers. *An Introduction to Machine Translation.* Academic Press, 1992.

[6] D.B. Jones. *The Processing of Natural Language by Analogy with Specific Reference to Machine Translation.* PhD thesis, UMIST, Manchester, 1991.

[7] J.L. McClelland and D.E. Rumelhart. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 2: Psychological and Biological Models.* MIT Press, 1986.

[8] I.J. McLean. *A Study of Recurrent Connectionist Architectures for Unsupervised Temporal Pattern Recognition.* Master's thesis, University of Manchester, 1991.

[9] M. Nagao. A Framework of Mechanical Translation between Japanese and English by Analogy Principle. In A. Elithorn, editor, *Artificial and Human Intelligence,* pages 173-180. Elsevier, 1984.

[10] D.E. Rumelhart and J.L. McClelland. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations.* MIT Press, 1986.

[11] V. Sadler. The textual knowledge bank: Design, construction and applications. In *International Workshop on Fundamental Research for the Future Generation of Natural Language Processing,* pages 17-32. ATR Telephony Laboratories, 1991.

[12] S. Sato and M. Nagao. Towards memory based translation. In *COLING 90* (Helsinki), volume 3, pages 247-252, 1990.

[13] R. Skousen. *Analogical Modeling of Language.* Kluwer Academic Publishers, 1989.

[14] E. Sumita, H. Iida, and H. Kohyama. Translating with Examples: A New Approach to Machine Translation. In *The Third Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages.* Linguistics Research Center, University of Texas at Austin, pages 203-212, 1990.