Context-Sensitive Retrieval for Example-Based Translation

Ralf D. Brown
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
ralf+@cs.cmu.edu

Abstract

Example-Based Machine Translation (EBMT) systems have typically operated on individual sentences without taking into account prior context. By adding a simple reweighting of retrieved fragments of training examples on the basis of whether the previous translation retrieved any fragments from examples within a small window of the current instance, translation performance is improved. A further improvement is seen by performing a similar reweighting when another fragment of the current input sentence was retrieved from the same training example. Together, a simple, straightforward implementation of these two factors results in an improvement on the order of 1.0–1.6% in the BLEU metric across multiple data sets in multiple languages.

1 Introduction

While context has long been recognized as an important factor in translating texts, it tends to be given lower priority in machine translation system development than improving the quality of isolated translations. Quality can only be improved so far, however, when operating strictly on isolated sentences, and thus further improvements must eventually be sought by taking other sentences into account when performing a translation.

EBMT systems typically treat both training data and the input to be translated as bags of unrelated sentences, though in practice, consecutive sentences are in fact related. Rather than consisting of random sentences, the training data consists of a set of coherent documents, and the input to be translated is one or more documents. In particular, retrieval is done without regard to the results of the prior sentence's translation, and thus differing word senses receive equal weighting. In contrast, by considering whether the previous sentence that was translated used adjacent sentences in the training corpus, the appropriate word sense can be given more import in the final translation, based on the old idea of "one sense per discourse" (Gale et al., 1992). A similar idea of temporal coherence in the use of word senses is used in speech recognition in the form of trigger or cache models for disambiguating homophones.

Figure 1 shows an example of using context to select the appropriate word sense for a translation. The training material includes examples for three senses of the word "bank", two of which produce equally-long matches between the training data and the second sentence of the test input. Without using context, the system can't distinguish between those two matches (which would generate "Ufer" and "Bank" in German, for example). However, by giving a bonus to the match where a nearby training instance was used in generating the first sentence's translation, the hypothesis with the correct "financial institution" sense can be given priority in generating the overall translation.

Similarly, for an EBMT system which uses partial matches of training examples (either explicitly partial matching as in (Brown, 1996; Brown, 2001; Brown, 2004) or complete matches of training instances which may be fragments of the original example sentences as in (Veale and Way, 1997; Gough and Way, 2003)), having multiple matches between the test input and a single training sentence increases confidence in the correctness of *all* matches in that sentence.

The next two sections of this paper describe the implementation of these two simple approaches to taking advantage of context.

2 Local Context

The EBMT system (Brown, 1996; Brown, 2004) used for the experiments described in this paper retrieves contiguous fragments from the training corpus which exactly match portions of the input to be

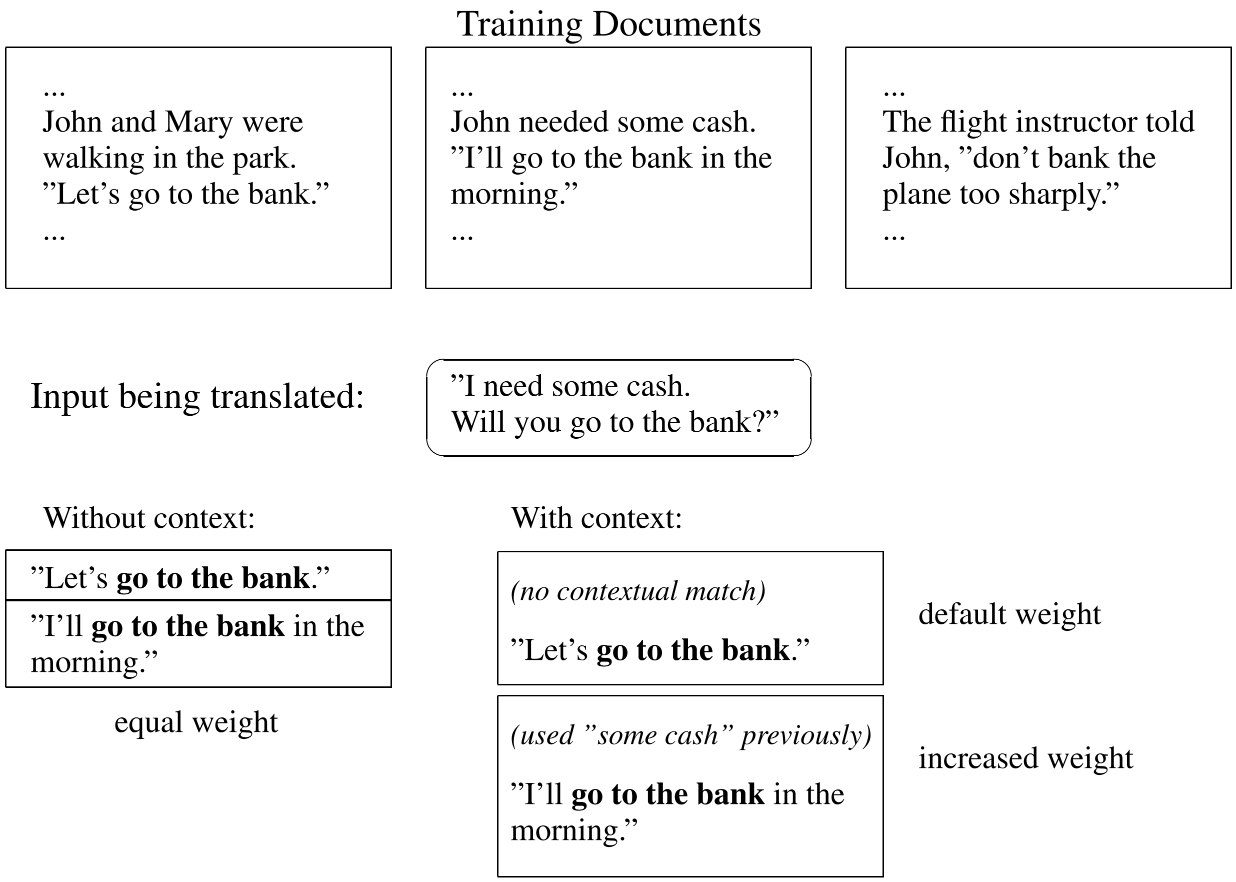

Figure 1: Adjacent sentences affect quality of the retrieved examples

translated[1]. Thus, if a new sentence is largely the same as a training example but contains a section which differs, two (or more) fragments will be retrieved from that example. Clearly, two fragments retrieved from a single example are better than the same fragments retrieved from two different examples (Figure 2). Thus, the translation hypothesis generated by a retrieved partial example should be given greater weight if other fragments of the input text occur in the same training example.

Further, since the system retrieves *every* phrasal match, whenever it finds e.g. a four-gram match, the trigrams and bigrams contained within it contribute to the pool of examples for determining the candidate translations of those trigrams and bigrams. However, the initial implementation did not take advantage of the fact that such contained instances are more reliable because they occur in an appropriate context, while n-gram instances which are not contained within a longer match do not have the same context as the phrase in the test input.

Thus, *local context* can guide the selection of appropriate translation hypotheses by boosting the weight given to a retrieved match whenever other matches of the current input sentence occur within the same training example. For ease of implementation, the initial version of local context weighting uses a greedy one-pass approach rather than separate passes to collect statistics and weight retrieved examples. As a result, some matches receive less of a boost than they should, but the overall impact is expected to be fairly small. By far the most frequent recipients of a bonus are bigrams contained within larger matches, but many of them are never actually processed because (for speed reasons) the EBMT system only examines up to a maximum number of matches for any particular n-gram of the input, typically 1000 or 1500.

Differential weighting based on the local con-

[1] Or exactly match all or a portion of a generalized template formed from the input, but that feature was not used for the experiments described here.

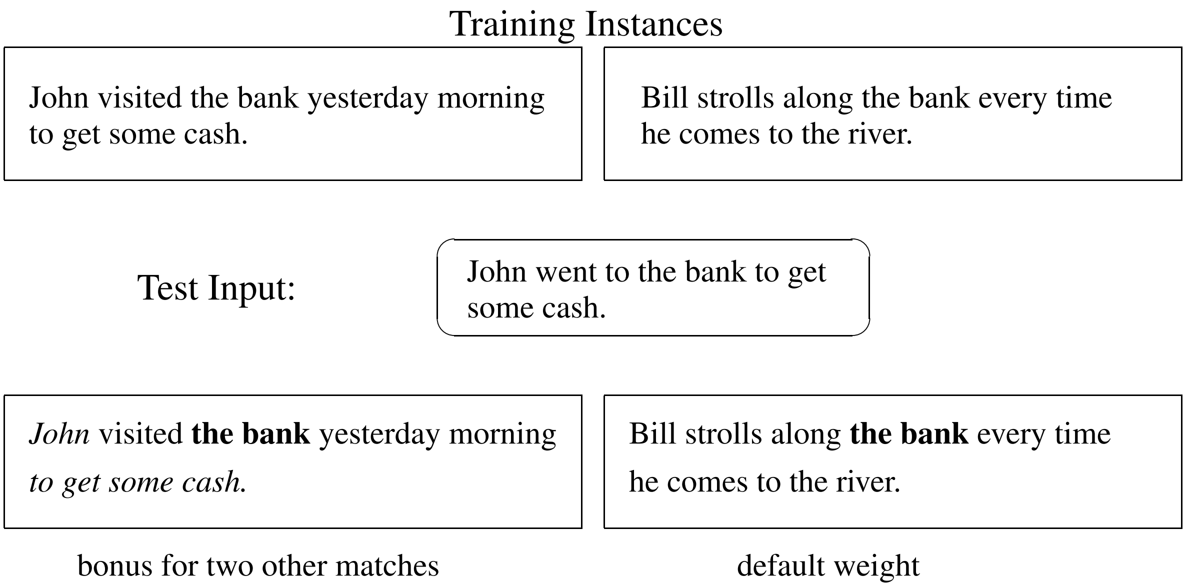

Figure 2: The quality of retrieved fragments varies by relative location.

text was implemented as an extension to an existing differential-weighting mechanism. Each retrieved instance receives a weight based on a combination of the source of the training data and its proportional location in the corpus. For example, when translating newswire texts, newswire training data could receive a weight of 3.0 and parliamentary proceedings a weight of 1.0; and when translating current texts using a corpus gathered over a long period of time, the earliest example could receive a weight of 1.0, linearly increasing to 2.0 for the most recent example in the corpus (all of these weights are configurable). When computing the confidence score for each distinct candidate translation, a weighted sum of all the retrieved instances is used to compute a translation probability, which forms the bulk of the quality score (the highest alignment-confidence score for any instance generating a particular translation forms the remainder of the score). Thus, increasing the weight of a training example increases the translation probability and hence the overall confidence score assigned to the associated translation. This causes a re-ranking of the translation hypotheses for a particular source phrase, and can result in a different set of hypotheses being output whenever there are more distinct translation hypotheses than the system has been configured to produce.

To compute the local context bonus assigned to a retrieved training instance, an array is used to keep counts of all retrievals from each training example in the corpus. The counts are initialized to zero and incremented each time a match from the associated training example is accessed. The base weight of the instance (as described in the previous paragraph) is multiplied by one plus a configurable bonus factor times the total access count. A fairly large bonus factor, typically on the order of 10, is required to counteract the sheer number of other matches which do not receive a bonus and thereby produce a substantive shift in the relative weighting of different translation alternatives. The matches found by examining the index are processed in order from longest to shortest, so a short match contained within a longer one automatically receives a local context bonus. Because a one-pass algorithm was implemented, only the second and subsequent disjoint fragments matching a given training instance will receive a bonus; the first fragment processed will not.

3 Inter-Sentential Context

As mentioned in the introduction, EBMT systems typically treat both training data and the input to be translated as bags of unrelated sentences. But in practice, consecutive sentences are in fact related – the training data consists of a set of coherent documents, and the input to be translated is one or more documents rather than random sentences.

Given the implementation of the local context

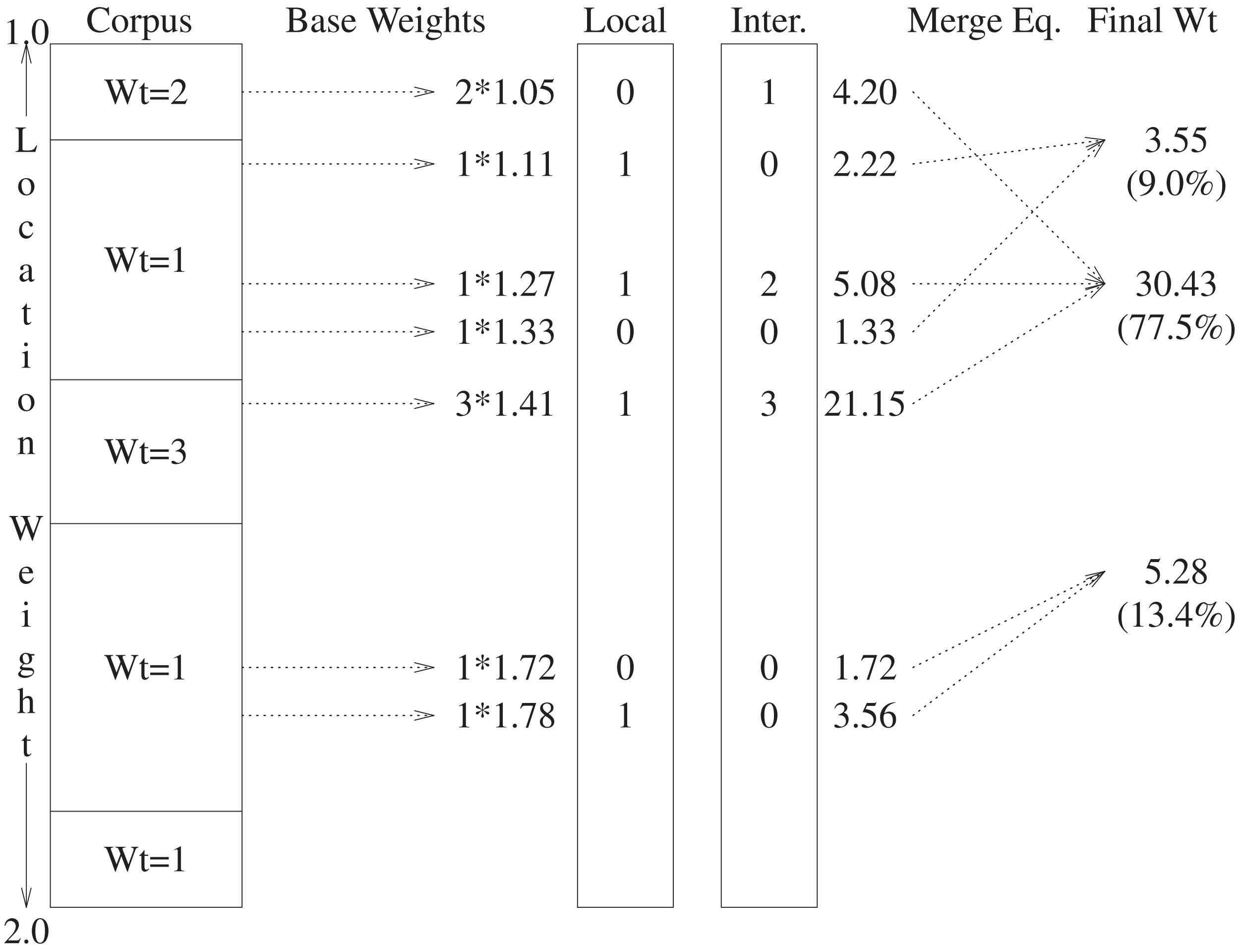

Figure 3: Computing weighted translation probabilites with context bonuses

mechanism described in the previous section, implementation of intersentential context bonuses is very simple: rather than discarding the usage counts after translating an input sentence, they are retained for the following translation, separately from the new local context counts. During the second sentence's translation, the counts within a selected range around each retrieved instance are consulted. The intersentential context bonus is then the weighted sum of the counts within the selected range (in the current implementation, the current example plus the five examples before and after it, though the most distant of those five examples generally receive zero weight).

For example, let the bonus weights be set to 10 for the current training example, 5 for the examples immediately adjacent, and 2 for the examples at distance two, e.g. (2 5 10 5 2). The total bonus for an example where the previous example had one match during the prior translation and the example two sentences later in the corpus had two matches would be $(1 * 5) + (2 * 2)$ or 9.

Intersentential context weights are factored into the base weight of a retrieved instance in the same manner as local context weights, making the final weight of each instance the product of its base weight times one plus the sum of its local context bonus and its intersentential context bonus.

The final weight of a translation alternative is the sum of the individual weights of each of the instances which generate that alternative, computed as just described. See Figure 3 for a visual representation of this process.

4 Experimental Design

To determine the efficacy of the two context bonuses, multiple test sets were translated and scored using the BLEU metric under each of four conditions:

- **baseline**: no context bonuses
- **local**: only local context bonus applied
- **intersent.**: only intersentential context bonus
- **both**: both bonuses applied

Each of the four conditions was separately tuned to determine the best values for several key parameters of the EBMT system (maximum number of hypotheses for a given source phrase, alignment confidence threshold, proportion of confidence score from translation probability, and relative importance of target-language trigram language model). The intent was to show the maximum performance possible for each context bonus and for the combination of the two bonuses to evaluate their potential benefit.

Four language pairs were used: French-English, Spanish-English, Chinese-English, and Romanian-English. For each language pair, two test sets were selected, one on which to tune (producing peak-to-peak comparisons between the experimental conditions), and one as held-out data to estimate real-world performance on unseen test data.

The French-English EBMT system was trained on 20,000 sentence pairs from files 000 and 001 of the IBM Hansard corpus (Linguistic Data Consortium, 1997). The test sets were 100 sentence pairs drawn from file 020 for tuning and 1000 sentence pairs drawn from file 060 for evaluation.

The Spanish-English system was trained on some 700,000 sentence pairs (approximately 22 million words) from the UN Multilingual Corpus, about one-tenth that amount of text from European Parliament proceedings, and a small amount of text from the Pan-American Health Organization. The test sets were 280 and 1389 sentences, respectively, held out from the European Parliament texts.

The Chinese-English system was trained on slightly less than two million sentence pairs drawn primarily from the UN Chinese-English corpus available from the Linguistic Data Consortium. The test sets were the 993-sentence test set from the 2002 DARPA TIDES Machine Translation Evaluation for tuning and the 919-sentence test set from the 2003 MT Evaluation as unseen data, both primarily newswire text.

The Romanian-English system was trained on the parallel corpus provided to participants in the shared word-alignment task for the 2003 and 2005 Workshops on Parallel Text (Mihalcea and Pedersen, 2003), approximately one million words per language. The 2003 test set of 248 sentences was used as the tuning set, and the 2005 test set of 203 sentences was intended for use as the unseen test data. Unfortunately, the latter set proved to consist of sentences drawn from the training corpus, which thus made it unusable without first modifying the training data to remove those sentence pairs (as the EBMT system produced perfect matches for the reference translations regardless of settings). Therefore, only one test set was used for Romanian-English experiments.

We performed significance tests on the experiments using the four test sets of around 1000 sentences (the other three test sets were too small to produce reliable results). To compute the statistical significance of changes in performance, the test set was split into ten approximately equal-sized parts and BLEU scores computed for each part. The two-tailed version of Student's paired t-test was applied to the sets of scores to compute p-values.

The BLEU metric uses a global brevity penalty to partially compensate for its lack of direct recall measurement. Because this penalty more easily becomes substantial with smaller test sets, the average score obtained on a set of smaller files tends to be somewhat lower than the score obtained on the concatenation of those files (where the natural variability in translation lengths tends to be smoothed out). The reduction averaged slightly more than 2 percent over the various combinations of test condition and test set on which the ten-way split was used.

5 Results

For all four language pairs, each of the two classes of context alone and in combination resulted in improved performance when pitted against the original implementation without context awareness (Table 1). The "real-world" performance on previously-unseen data using the optimal parameters determined on the tuning set was rather mixed (Table 2) for intersentential context and the combination of local and intersentential, but local context still provided a statistically significant improvement in two of three cases (statistically-significant differences are shown in boldface in Tables 1 and 2).

Three of the four larger test sets for which significance could be computed achieved statistically significant improvements in BLEU scores. For Spanish-English, there was extremely high variance between the ten slices of the test set (in particular, one slice scored less than half the average, possibly

Language	Test Size	Local	Intersent.	Both
French	100	+0.71%	+0.97%	+1.03%
Chinese	993	+1.36%	+0.58%	**+1.69%**
Romanian	248	+0.86%	+0.79%	+1.44%
Spanish	280	+1.36%	+0.63%	+1.36%

Table 1: Relative Improvements from Using Context (Peak-to-Peak)

Language	Test Size	Local	Intersent.	Both
French	1000	**+1.51%**	+0.33%	-0.26%
Chinese	919	**+0.83%**	-0.33%	+1.08%
Spanish	1389	+1.22%	-0.60%	-0.28%

Table 2: Relative Improvements from Using Context (Unseen Test Data)

due to errors or divergences[2] in the available translation), and thus resulted in a non-significant p-value of 0.20 even for local context.

6 Conclusions

Although very simple, the implementation of local context described in this paper proves to be beneficial in all cases, while the simple implementation of intersentential context is more of a mixed bag in terms of performance. The computation of intersentential context bonuses is probably being affected by document boundaries, which are not being taken into account. Particularly where the original documents are short, such as newswire stories, even a three-sentence window on either side of the current instance has a good chance of including text from another document.

Because the contextual bonuses result in a reranking of hypotheses, it is possible for the local and intersentential bonuses to act against each other. This is likely what happened on the larger French test set, where the two bonuses individually produced improvements in the BLEU score while the combination was actually detrimental.

It is interesting to note that the only language pair on which the combination of local and intersentential contexts improved performance on the unseen data is also the only language pair where the tuning set was itself large enough to perform statistical significance tests. The failure to produce an improvement may therefore simply be a result of tuning sets which were too small to find appropriate parameter settings for the general case, rather than just the limited number of sentences used for tuning.

7 Future Work

As a first, very quick implementation, many enhancements still await implementation and investigation. Two enhancements which have already been mentioned are two-pass calculation of bonuses and consideration of document boundaries. Other, more global, matching is also likely to improve performance.

Two-pass calculation of contextual bonuses will eliminate the cases where the existing one-pass calculation does not give a retrieved instance as much of a context bonus as it should receive, because not all of the contextual instances which contribute to the bonus have been processed yet. For intersentential context, using two passes in a batch mode will also permit the assignment of a bonus based on following sentences in the input, e.g. if the input sentences in Figure 1 were reversed, the appropriate sense of “bank” would still receive a bonus. Naturally, some applications of machine translation require production of a translation immediately upon receipt of a sentence, and in those applications such batching will not be possible (but a two-pass calculation can still be used for local context).

Consideration of document boundaries will eliminate the cases where a sentence from another document contributes to the intersentential context bonus merely because it lies within the window being considered.

Finally, where the fine-grained document boundaries are available, the base weights assigned to re-

[2] In at least one case, two consecutive sentences were translated with some of the information from one moved to the other in the translation.

trieved matches can be dynamically adjusted. When performing a batch translation of a document, a global similarity can be computed between the input document and each of the training documents, and base weights adjusted upwards for the most similar documents. This then automatically biases the translations towards those used in the documents which are most similar in subject matter, style, and genre to the input text, much as the current code permits a static adjustment of weights by the user to match the anticipated domain of the text to be translated.

Orthogonal to all of the above enhancements, more investigation is needed to ensure that improved scores on the tuning data reliably result in improved scores on unseen texts.

References

Ralf D. Brown. 1996. Example-Based Machine Translation in the PANGLOSS System. In *Proceedings of the Sixteenth International Conference on Computational Linguistics*, pages 169–174, Copenhagen, Denmark. `http://www.cs.cmu.edu/~ralf/papers.html`.

Ralf D. Brown. 2001. Transfer-Rule Induction for Example-Based Translation. In *Proceedings of the Workshop on Example-Based Machine Translation*, September. `http://www.cs.cmu.edu/~ralf/papers.html`.

Ralf D. Brown. 2004. A Modified Burrows-Wheeler Transform for Highly-Scalable Example-Based Translation. In *Machine Translation: From Real Users to Research, Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA-2004)*, volume 3265 of *Lecture Notes in Artificial Intelligence*, pages 27–36. Springer Verlag, September-October. `http://www.cs.cmu.edu/~ralf/papers.html`.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. One Sense Per Discourse. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York*, pages 233–237, February. `http://www.ldc.upenn.edu/H/H92/`.

Nano Gough and Andy Way. 2003. Controlled Generation in Example-Based Machine Translation. In *Proceedings of the Ninth Machine Translation Summit (MT Summit IX)*, pages 133–140.

Linguistic Data Consortium. 1997. *Hansard Corpus of Parallel English and French*. Linguistic Data Consortium, December. `http://www.ldc.upenn.edu/`.

Rada Mihalcea and Ted Pedersen. 2003. An Evaluation Exercise for Word Alignment. In *Proceedings of the HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 1–10. Association for Computational Linguistics, May.

Tony Veale and Andy Way. 1997. Gaijin: A Template-Driven Bootstrapping Approach to Example-Based Machine Translation. In *Proceedings of the NeMNLP'97, New Methods in Natural Language Processessing*, Sofia, Bulgaria, September. `http://www.compapp.dcu.ie/~tonyv/papers/gaijin.html`.