Machine Translation without a Bilingual Dictionary

Abstract

This paper outlines experiments conducted to determine the contribution of the traditional bilingual dictionary in the automatic alignment process to learn translation patterns, and at runtime. We found that by using automatically derived translation word pairs combined with a function word only lexicon, we were able to either match or nearly match the translation quality of the system that used a full traditional bilingual lexicon in addition. The language pairs studied were French-English and Spanish-English.

1. Introduction[1]

Bilingual dictionaries can be a curse as much as a blessing, especially when used automatically. Words in one language can map to many translations in another language, and getting the right translation in a certain context is not guaranteed, since the correct translation for the given context might not be in the dictionary. In addition, the translation provided by the dictionary may be inappropriate, because it does not fit the context. For example, we found that the word ***serveur*** in French was systematically translated as ***waiter*** (instead of ***server***) by a commercial system which depends heavily on bilingual domain dictionaries which are not sensitive to context.

Moreover, building bilingual dictionaries is very costly, and dictionaries need to be tailored to particular domains and kept up to date. Therefore, we are interested in the question of whether bilingual dictionaries are necessary when translation patterns can be learned automatically. This paper describes experiments which show that removing the bilingual dictionary from our system does not affect translation quality.

1.1 System overview

We review here the basics of the MSR-MT translation system, but refer the reader to Pinkham et al. (2001) and Richardson et al. (2001) for full details on the French and Spanish component creation. The architecture and components are the same for both systems.

MSR-MT uses broad coverage analyzers, a large multi-purpose source language dictionary, a large bilingual lexicon, an application independent English natural language generation component and a transfer component.

The transfer component consists of transfer patterns automatically acquired from sentence-aligned bilingual corpora using an alignment grammar and algorithm described in detail in Menezes & Richardson (2001). Training takes place on aligned sentences which have been analyzed by the source language and English analysis systems to yield dependency structures specific to our system entitled Logical Forms. The Logical Form structures, when aligned, allow the extraction of lexical and structural translation correspondences which are stored for use at runtime in the transfer database. The transfer database can also be thought of as an example-base of conceptual structure representations. See Figure 1 for an illustration of the training process.

[1] Many thanks to Robert Moore for suggesting that we experiment without the traditional bilingual dictionary; to Joseph Pentheroudakis for allowing us to conveniently work with function words only, and to Mike Carlson for extensive help in carrying out these complex experiments.

The transfer database is trained on more than 200,000 pairs[2] of aligned sentences from computer manuals and help files. Alignment relies on a general purpose bilingual dictionary, except when deliberately excluded, as in the experiment below. To make domain specific the vocabulary available to the alignment process, we add translation pairs extracted from the specific domain, using statistical word/phrase assignment. This results in French-English (FE) and Spanish-English (SE) files of automatically created translation correspondences, or word associations. These files, of approximately 30,000 word pairs, add to the quality of the alignments and to overall translation quality.

Furthermore, we will show in this paper that the word association file is sufficient to learn good translation patterns in the alignment process when supplemented with function word and stop-word translations, and that the bilingual dictionary is not necessary. For full details on the word association list creation, see Moore (2001) and Pinkham & Corston-Oliver (2001).

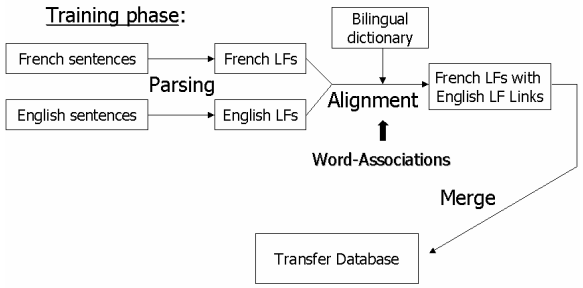

Figure 1

The word association file is used only in training (see Figure 1) to enhance the opportunity for alignment during the detection of transfer patterns. The bilingual dictionary was typically used in training and at runtime, except when deliberately excluded.

1.2 Experiments

The question that we address in this paper arose from earlier work on the French-English (FE) system. We noted that, when assembling the components of the system, we did not get a significant improvement with the addition of the learned word-association file. It became important to know why this was the case, so we modified the system to allow it to use the word-association file only at training time, and a function word dictionary both at training and at runtime[3]. This gave us a system without a bilingual dictionary as depicted in Figure 2. The previous system uses both the bilingual dictionary and the word association file at training and runtime, as in Figure 3.

[2] The French-English system has a training set of approximately 200,000 and the Spanish-English system uses 350,000 at this time

[3] Function words are defined as the class of words that are Pronouns, Conjunctions and Prepositions in our dictionary. For Spanish-English, the number of function words was approximately 520.

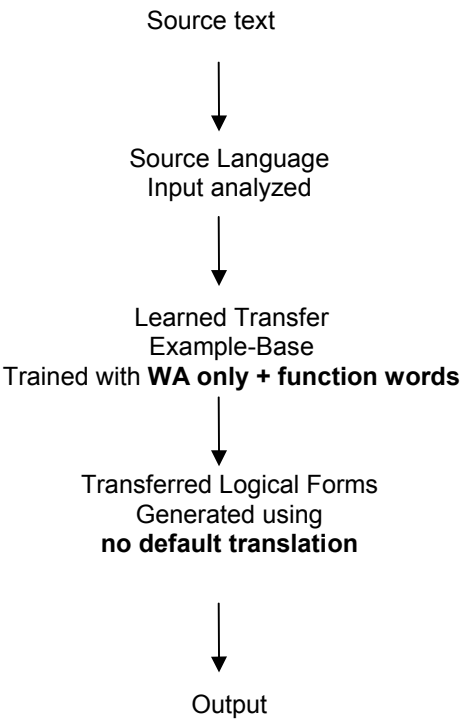

Figure 2

System without Bilingual Dictionary
WA = Word Association pairs

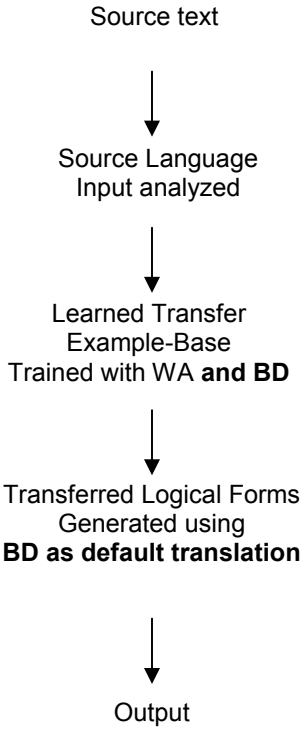

Figure 3

System with Bilingual Dictionary
BD = Bilingual Dictionary

Because the system without bilingual dictionary currently does not have a translation dictionary in which to look up translations at runtime, it will be obtaining all the translation information from the Transfer Example-Base. When there is no translation available, a French or Spanish word will appear in the English output. This is rare in the case of the system with bilingual dictionary, because of the large size of the Bilingual dictionaries for FE and SE, but does occur in approximately 15% of the sentences when the bilingual dictionary is not used. The experimental results presented in section 3 show that for French-English, there was no statistically significant difference in the quality of the output of the system-without bilingual dictionary and the system-with bilingual dictionary, where they were measured and scored against an outside metric, the commercial system Systran.

This result was very surprising to us. Our intuitions were that general purpose bilingual dictionaries should be of use in our system, particularly when seeding the alignment learning process. Because of the relative immaturity of the French-English system, we decided to conduct a parallel evaluation for the Spanish-English system, with a system-with (bilingual dictionary) and a system-without (bilingual dictionary), rating them against the benchmark used in previous experiments (Babelfish website). Full results are presented in section 3.

2. Experiment and Methodology

We performed several evaluations of machine translation quality, in French-English and Spanish-English, each with and without the bilingual dictionary rated against the benchmark. These evaluations were performed by an independent organization that provides support for Natural Language application development; the evaluators are completely independent of development activities.

2.1 Evaluation design

For each condition to be tested, seven evaluators were asked to evaluate the same set of 250 blind test sentences. For each sentence, raters were presented with a reference sentence, the original English translation from which the human French translation was derived. In order to maintain consistency among raters who may have different levels of fluency in the source language, raters were not shown the original French or Spanish sentence (for similar methodologies, see Ringger et al., 2001; White et al., 1993). Raters were also shown two machine translations, one from the system with the component being tested (System 1), and one from the comparison system (System 2). Because the order of the two machine translation sentences was randomized on each sentence, evaluators could not determine which sentence was from System 1. The order of presentation of sentences was also randomized for each rater in order to eliminate any ordering effect.

The raters were asked to make a three-way choice. For each sentence, the raters were to determine which of the two automatically translated sentences was the better translation of the (unseen) source sentence, assuming that the reference sentence was a perfect translation, with the option of choosing "neither" if the differences were negligible. Raters were instructed to use their best judgment about the relative importance of fluency/style and accuracy/content preservation. We chose to use this simple three-way scale in order to avoid making any a priori judgments about the relative importance of these parameters for subjective judgments of quality. The three-way scale also allows sentences to be rated on the same scale, regardless of whether the differences between output from System 1 and System 2 were substantial or relatively small; and regardless of whether either version of the system produced an adequate translation.

The scoring system is similarly simple; each judgment by a rater was represented as 1 (sentence from System 1 judged better), 0 (neither sentence judged better), or -1 (System 2 judged better). The score for each condition is the mean of the scores of all sentences for all raters.

3. Test results

The versions of our systems with and without bilingual dictionary were compared to an outside commercial system on both the FE system of May 2001 and the SE system of October 2001. As demonstrated by the negative score for our FE system, the quality of French-English at that time was considerably less good than the quality of the SE system in October. This is not at issue directly, since the French-English system is 9-12 months behind in terms of development in the machine translation context. As determined by a one-tailed t-test with a value of p=0.3, in a less mature system, the impact of the bilingual dictionary is not significant (see Figure 4).

Condition	Score	Significance	Sample size
Bilingual Dict	-.14 +/- .11	0.994	250
No Bilingual Dict	-.124 +/- .07	0.98	250

Figure 4: French-English (FE) results

Condition	Score	Significance	Sample size
Bilingual Dict	.43 +/- .10	>.99999	250
No Bilingual Dict	.34 +/- .12	>.99999	250

Figure 5: Spanish-English (SE) results

In the case of Spanish-English, there is a statistical significant difference between the better bilingual dictionary version and the no bilingual dictionary version, as determined by p= .004 in a one-tailed t-test. Note, however, that they are both considerably better than the benchmark, so the loss of quality is minor. (see Figure 5)

	NO-BD	BD
Sent. with diffs (61)	**-0.25246**	**0.161475**
Sent. without diffs (189)	**0.538624**	**0.521693**

Figure 6: Average scores for translations which differ with and without bilingual dictionary

When we break down the scoring into sentences where both versions of the SE system gave the same translation and those where they gave different translations, we can observe that sentences which accessed the bilingual dictionary at runtime (or in the case of the version without bilingual dictionary, had foreign words in them), were on the average considerably worse than sentences whose content came entirely from the Transfer Example-Base. Figure 6 gives the scores of the sentences for which the two versions of our system give different translations (1st row) and the sentences for which both versions give the same translation (2nd row). The scores are slightly different in this last case because the two evaluations were conducted by different raters. We conclude from this breakdown of scores that sentences that differed in the two systems were considered to be poor translations in both cases, an interesting discovery that we explore further in the next section.

4. Discussion

4.1 FE Evaluations

In the FE evaluations, the bilingual dictionary seemed to be redundant, because there was no significant difference whether it was used or not. However, the FE system at that time was in its early stages, and more than 50% of the translations with or without bilingual dictionary were judged inferior to the benchmark's translations. Whether or not source-language words occurred in the translation didn't make much of a difference in the rating, if the source sentence was badly analyzed in the first place: compared to Systran's translation, the score would be negative in any case.

For example, in the following sentence, both versions (with and without bilingual dictionary) get a score of -1 as compared to Systran because the translation is bad whether or not the word *temps* is translated. There is no appropriate treatment of verbal idioms *prendre du temps* ('take some time') and *il s'agit de* ('it concerns'). *SCE* introduces the source sentence, *REF* the reference, *BAB* the translation by Babelfish, *NO-BD* the translation without bilingual dictionary and *BD* the translation with the bilingual dictionary.

SCE: Ceci risque de prendre du temps, surtout lorsqu'il s'agit de lecteurs de disquette.
REF: This can take extra time, especially for floppy disk drives.
SYS: This is likely to take time, especially when it acts diskette drives.
NO-BD: This expects to support **temps**, especially when it acts as disk drives itself.
BD: This expects to support **a time**, especially when it acts as disk drives itself.

Under these conditions, it is impossible to draw conclusions about the usefulness of the BD, leading us to conduct the same experiment on a more mature system, the SE system[4].

4.2 SE Evaluations

Out of the 250 sentences used in the experiments, 61 were translated differently by the two versions of our system. The version with the SE bilingual dictionary was preferred in both sets of evaluation (between the two versions of the system, and against the commercial benchmark). However, the system without a bilingual dictionary was still much preferred over the commercial benchmark.

There are two types of differences in translation: in 45 out of 61 sentences, there are one or two Spanish words in the English translation without bilingual dictionary. In the remaining 16 sentences, the translations differ in their formulation, even though no source-language words appear.

4.2.1 No source-language words in the translation

Even when there are no Spanish words in the translation without the bilingual dictionary, there may be differences between the translations of the two versions of the system: the bilingual dictionary is used by the transfer component of the version with bilingual dictionary to establish links between aligned corpora. Those links can differ if the dictionary is not used during training, and this explains the differences in translation. Out of the 16 sentences which differ in their formulation, 11 translations are as good or better (9 better) when the bilingual dictionary is not used and 5 translations are worse. The average score for these translations against Babelfish is .46 without the bilingual dictionary, and .2 with the bilingual dictionary. This means that the use of the dictionary results in the creation of bad links, or prevents the creation of good links. A couple of examples of translations are given below.

In the translation without bilingual dictionary, a link is established between *llevar a cabo* and 'conduct'. In the translation with bilingual dictionary, there is no such link, and the translation of both words comes from the bilingual dictionary: 'take' for *llevar*, 'to corporal' for *a cabo*.

SCE: Cuando realice una instalación nueva de Windows 2000 Server, podrá seleccionar la partición en la que desea llevar a cabo la instalación.
REF: When you perform a new installation of Windows 2000 Server, you can select the partition on which to install.
BAB: When it makes a new installation of Windows 2000 Server, it will be able to select the partition in which it wishes to carry out the installation.
NO-BD: When you perform a new installation of Windows 2000 Server, you will be able to select the partition in which you want to ***conduct*** the installation.
BD: When you perform a new installation of Windows 2000 Server, you will be able to select the partition in which you want to ***take*** the installation ***to corporal***.

In the following sentence, there is a link in the system without bilingual dictionary between *archivo de registro de importación de base de datos* and 'database-import log file'. In the system with bilingual dictionary, however, this link is absent, and the translation of the expression is less fluent (there are links, however, between *archivo de registro* and 'log file',

[4] We decided to use SE, because its quality as reported in Richardson et al. (2001) surpasses the competitor.

and between *importación de base de datos* and 'database-import', and the translation is quite acceptable also).

SCE: Los siguientes ejemplos son entradas (solicitud de acceso y aceptación de acceso) de un archivo de registro de importación de base de datos.
REF: The following are sample entries (access-request and access-accept) from a database-import log file.
BAB: The following examples are entered (request of access and acceptance of access) of a file of registry of import of data base.
NO-BD: The following examples are entries Access-Request and acceptance of access in a ***database-import log file***.
BD: The following examples are entries Access-Request and acceptance of access in a ***log file of database-import***.

4.2.2 Source language words in the translation

Most of the differences in translation between the two versions of the system come from the presence of Spanish words in the English translation. At run-time, the bilingual dictionary is used when words of the source are not in the transfer database. If these words are not in the bilingual dictionary either, or if the bilingual dictionary is not used, they occur non-translated in the translation.

There are three subclasses to distinguish: cases where there are problems in the linguistic analysis of the source sentence; cases where the source-language word is part of a stop list of words excluded from the alignment process; cases where the only problem of the translation is the presence of the Spanish word.

4.2.2.1 Problems in analysis

The first thing to notice is that Spanish words occur when there are problems in the analysis of the input sentence (in 14 out of 45 sentences). For example, in the sentence below, *abajo* is analyzed as a verb instead of a preposition, and this creates a bad analysis. The word ends up not being translated in either version, and the translation is quite bad also as a result of the wrong analysis: the purpose clause ends at *tops,* and the main clause starts with *I* (*abajo* is a first person singular).

Had the preposition been analyzed as such, it would have been found in the dictionary, and in the function word list for the version with bilingual dictionary. Also, the analysis would certainly have been better.

SCE: Para asegurarse de que el marco seleccionado se mueve hacia arriba o hacia abajo junto con el párrafo al que está fijado, active la casilla de verificación Mover con el texto.
REF: To ensure that the selected frame moves up or down with the paragraph it's anchored to, select the Move with text check box.
BAB: In order to make sure that the selected frame moves upwards or downwards along with the paragraph to which is fixed, it activates the square of verification To move with the text.
NO-BD: To ensure that the selected frame is moved towards tops, I ***abajar*** along with the paragraph estar to it pinned , select the Move check box with the text to it.
BD: To ensure that the selected frame is moved towards tops, I ***abajar*** along with the paragraph it is pinned to it , select the Move check box with the text to it.

In the next example, there are Spanish words only in the version without bilingual dictionary. However, the version with a dictionary is not significantly better. The verb *cancela* (*cancel*) is analyzed as a noun, which has a very specialized meaning: *lattice gate*. The conjunction *mientras* (*while*) is analyzed as an adverb and not as a conjunction introducing the verb *esta mostrando*, and this also concurs to the bad analysis. The version with bilingual dictionary has a higher score than the one without a bilingual dictionary, but not significantly better (-1 (no bilingual dictionary) / -.75 (bilingual dictionary)).

SCE: Si cancela la instalación del Adaptador de Acceso telefónico a redes mientras se está mostrando el cuadro de diálogo Copiando archivos, la instalación continuará y aparecerá el siguiente mensaje:

REF: If you cancel the Dial-Up Adapter installation while the Copying Files dialog box is displayed, the installation proceeds and returns the following message:

BAB: If it cancels the installation of the Adapter of telephone Access to networks while one is being to the dialogue panel Copying archives, the installation will continue and appear the following message:

NO-BD: If the installation of the Dial-Up Adapter is displaying the dialog box ***mientras cancela***, you will continue, Copying files, the installation, and the message will appear the following message Copies them:

BD: If the installation of the Dial-Up Adapter is displaying the dialog box while ***lattice gate***, you will continue, Copying files, the installation, and the message will appear the following message Copies them:

4.2.2.2 Sentences with stop-words

Stop-words are words which are excluded explicitly from alignment on their own. They only occur in links which involve other words. They include verbs which can act as "light verbs", for example, which do not have much independent semantic content, but acquire content in context. An example is *make*: its meaning differs in "*make someone do something*" and "*make a blunder*". Some words which are not light verbs are also explicitly excluded from the alignment process when they are not part of a larger unit: for example *cosa ('thing')*, *persona ('people')*, *lugar ('place')*, *parte ('part')*.

These words are excluded from the alignment process in isolation, because they are so heavily dependent on context. They only occur in links which also include some context. When they occur at runtime in a context which has not been encountered during training, their translation always comes from the bilingual dictionary[5].
Because the translation of these verbs in isolation can never be learnt during training, they are sometimes not translated when the bilingual dictionary is not used (if they happen not to be part of a larger link, i.e. a link encompassing several words).

Altogether, out of 45 translations with Spanish words, there are 32 sentences with one or more of these stop-words. Apart from the presence of Spanish words, the translations of these sentences are usually good, with only 9 sentences exhibiting problems of linguistic analysis. The System without bilingual dictionary loses unfairly here against the system with a bilingual dictionary. A well-analyzed sentence is usually well translated, but the translation with a remaining source-language word will systematically have a lower score than a good

[5] These words in isolation are excluded from alignment because otherwise, bad translations are systematically learnt.

translation with no source-language word. Below are some examples of translations with this type of word.

Translations with and without the bilingual dictionary are identical, except for the Spanish word in the translation without bilingual dictionary, and the scores of the translations with bilingual dictionary are much higher when compared to Babelfish: -0.2 and -1 for the translation without bilingual dictionary and +0.6 and +1 with the bilingual dictionary.

SCE: Si tiene un árbol y un contexto predeterminados, una vez que ha iniciado una sesión no necesita volver a iniciar otra sesión o proporcionar otra contraseña para tener acceso a cualquier volumen del árbol predeterminado.
REF: If you have a default tree and context, once you have logged on you do not need to log on again or supply another password to access any volume in your default tree.
BAB: If it has a predetermined tree and a context, once it has initiated a session does not need to return to initiate another session or to provide another password to have access to any volume of the predetermined tree.
NO-BD: If it ***tener*** a default tree and context as soon as it has started, a session does not need to start another session again or again provide another password to have access to any volume of the default tree.
BD: If it ***has*** a default tree and context as soon as it has started, a session does not need to start another session again or again provide another password to have access to any volume of the default tree.

SCE: Hay cuatro modos de utilización de datos de Microsoft Access con Microsoft Word:
REF: There are four ways you can use Microsoft Access data with Microsoft Word:
BAB: There are four ways of use of data of Microsoft Access with Microsoft Word:
NO-BD: Four modes of using Microsoft Access data with Microsoft Word ***Habers***.
BD: ***There are*** four modes of using Microsoft Access data with Microsoft Word.

Our intuition is that the exclusion of stop-words from the translation database was responsible for the worst performance of the version without bilingual dictionary. We hypothesized that we could correct these by adding stop-word translations to our function-word list, thus avoiding these bad scores. To determine the impact of such an experiment, we recomputed scores leaving out all instances of sentences with stop-word issues.

In the table below, the class "light verbs" refers to the set of sentences with light verbs which are not translated in the version without bilingual dictionary[6]. The "others" set is the rest of the sentences with different translations with and without bilingual dictionary. This table shows that if the sentences with light verbs are excluded, the difference in score between the System with bilingual dictionary and the System without bilingual dictionary is no longer significant.

	BD	*NO-BD*	*DIFF*
Light verbs	0.285714	-0.56190476	0.847619
Others	0.09625	-0.09	0.18625

[6] This group of verbs is very small in number (9): examples are "hacer"(make), "tener" (have).

4.2.2.3 Other sentences with Spanish words

Finally, out of the 45 sentences with Spanish words, only 8 sentences have no problem other than the presence of the source word. There are cases where translations are not found during training because of scarce data. An example is given below. Such sentences obtain a good score with bilingual dictionary (0.8), but the lowest score without (-1).

SCE: Las flechas azules comenzarán a destellar sobre el tablero.
REF: Blue arrows will then begin flashing on the table.
BAB: The blue arrows will begin to flash on the board
NO-BD: The blue arrows will begin ***destellar*** about the table.
BD: The blue arrows will begin ***flashing*** about the table.

5. Conclusion and future work

These results show that the bilingual dictionary does not play an important role in translation with our system. Other factors account for the differences in score between the versions with and without bilingual dictionary.

Scores are systematically worse when source-language words are not translated, and this is due to three causes: stop-words, problems in linguistic analysis, and scarce data. Of these three factors, stop-words have the highest impact. We note that when these cases are excluded, the difference in performance between the two systems is no longer significant. In fact, for translations which do not have source-language word issues, the overall performance of the System without bilingual dictionary is better.

These results motivate us to experiment with not using a general bilingual dictionary at all. We are planning two types of experiments. In the first one, the word-association file, which is in fact a learned bilingual lexicon, will be modified to become a runtime bilingual dictionary containing part-of-speech information. We verified by hand that words missing in the translations without bilingual dictionary were indeed in the word association file (32 words), so we are confident that the modified version will contain the necessary information to replace the bilingual dictionary.

In our second type of experiments, stop-words will be included in the function-word list to balance their exclusion from the alignment process when in isolation.

Another set of experiments will be performed on the more mature FE system, which now performs better than Systran according to the latest independent evaluations. We will again evaluate the performance of our system with and without bilingual dictionary, with the same conditions as for the SE system.

We expect to further demonstrate that a general purpose bilingual dictionary is not necessary when the transfer database is learned by training on domain corpora, and that it can even be detrimental to the quality of the translation.

References

Frederking, Robert, and Ralf Brown. 1996. The Pangloss-Lite Machine Translation System. In Proceedings of the Conference of the Association for Machine Translation in the Americas. 268-272.

Melamed, I. Dan. 1996. Automatic Construction of Clean Broad-Coverage Translation Lexicons. In Proceedings of the Second Conference of the Association for Machine Translation in the Americas. 125-134.

Menezes, Arul and Steve Richardson. 2001. A Best-First Alignment Algorithm for Automatic Extraction of Transfer Mappings from Bilingual Corpora. In Proceedings of the Data-Driven MT workshop, ACL 2001.

Moore, Robert C. 2001. Towards a Simple and Accurate Statistical Approach to Learning Translation Relationships between Words. In Proceedings of the Data-Driven MT workshop, ACL 2001.

Pinkham, Jessie and Monica Corston-Oliver. 2001. Adding Domain Specificity to an MT system. In Proceedings of the Data-Driven MT workshop, ACL 2001.

Pinkham, Jessie, Monica Corston-Oliver, Martine Smets and Martine Pettenaro, 2001. Rapid assembly of a large-scale French-English MT system. In Proceedings of the 2001 MT Summit.

Richardson, Stephen, William B. Dolan, Arul Menezes and Jessie Pinkham. 2001. Achieving commercial-quality translation with example-based methods. In Proceedings of the 2001 MT Summit.

Ringger, Eric K., Monica Corston-Oliver, and Robert C. Moore. 2001. Using Word-Perplexity for Automatic Evaluation of Machine Translation. Unpublished ms.

Hideo Watanabe, Sadao Kurohashi and Eiji Aramaki. 2000. Finding Structural Correspondences from Bilingual Parsed Corpus for Corpus-based Translation. In Proceedings of COLING: The 18th International Conference on Computational Linguistics. 906-912.

White, John S., Theresa A. O'Connell, and Lynn M. Carlson. 1993. Evaluation of machine translation. In Human Language Technology: Proceedings of a Workshop (ARPA). 206-210.