Adding Domain Specificity to an MT system

Jessie Pinkham
Microsoft Research
One Microsoft Way
Redmond, WA 98152
jessiep@microsoft.com

Monica Corston-Oliver
Butler Hill Group
moco@butlerhill.com

Abstract

In the development of a machine translation system, one important issue is being able to adapt to a specific domain without requiring time-consuming lexical work. We have experimented with using a statistical word-alignment algorithm to derive word association pairs (French-English) that complement an existing multi-purpose bilingual dictionary. This word association information is added to the system at the time of the automatic creation of our translation pattern database, thereby making this database more domain specific. This technique significantly improves the overall quality of translation, as measured in an independent blind evaluation.

1 Introduction

The machine translation system described here is a French-English translation system which uses a French broad coverage analyzer, a large multi-purpose French dictionary, a large French-English bilingual lexicon, an application independent English natural language generation component and a transfer component. The transfer component consists of high-quality transfer patterns automatically acquired from sentence-aligned bilingual corpora using an alignment grammar and algorithm described in detail in Menezes (2001) (see Figure 1 for an overview of the French-English MT system).

The transfer component consists only of correspondences learned during the alignment process. Training takes place on aligned sentences which have been analyzed by the French and English analysis systems to yield dependency structures specific to our system entitled Logical Forms (LF). The LF 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 2 for an illustration of the training process.

The transfer database for French-English was trained on approximately 200,000 pairs of aligned sentences from computer manuals and help files. In these aligned pairs, the French text was produced by human translators from the original English version.

Sample sentences from the training set are:

French training sentence:
Dans le menu Démarrer, pointez sur Programmes, sur Outils d'administration (commun), puis cliquez sur Gestionnaire des utilisateurs pour les domaines.
English training sentence:
On the Start menu, point to Programs, point to Administrative Tools (Common), and then click User Manager for Domains.

The French-English lexicon is used during the training period of the transfer component to establish initial, tentative, word correspondences during the alignment process. The sources for the bilingual dictionary were: Cambridge University Press English-French, Soft-Art

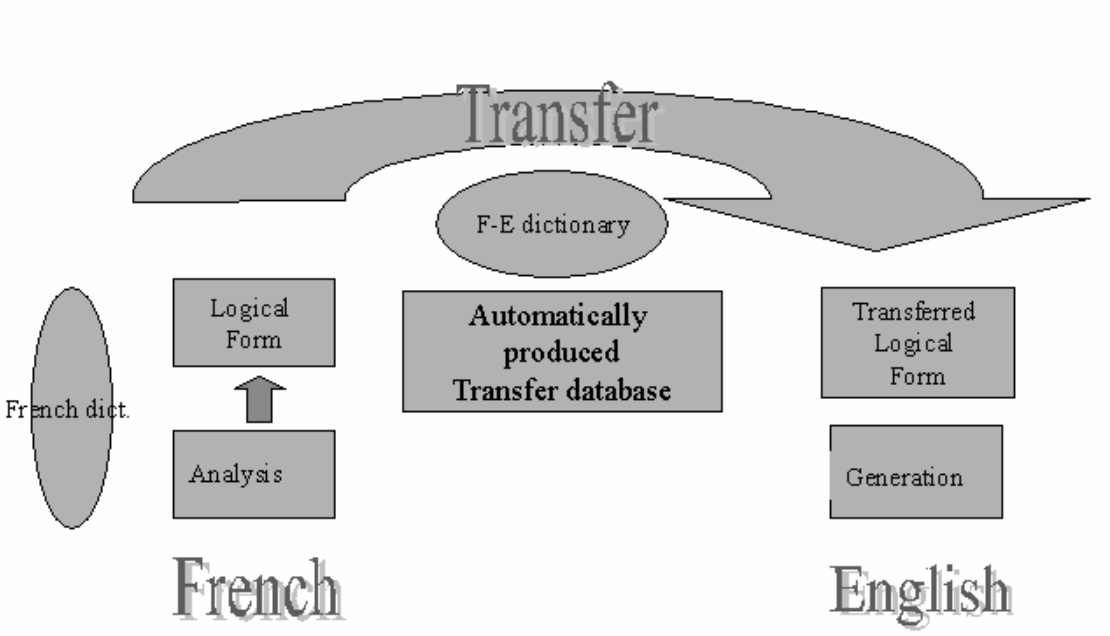

Figure 1

English-French, and Langenscheidt French-English and English-French dictionaries. The English-French translation data was reversed to create French-English pairs in order to augment the size of the dictionary, with a final translation count of 75,000 pairs.

However, quick examination of the sample sentence above shows that many terms are highly specific to the domain, e.g *menu Démarrer* <-> *Start menu*. To further add to the specificity of the vocabulary available to the alignment process, we added translation pairs extracted from the actual domain, using statistical word/phrase assignment, as described below. This resulted in one file of automatically created French English translation correspondences, or word associations (WA), and a second file of specialized multi-word translation correspondences which we term Title Associations (TA). These files, of size 30,000 and 2600 respectively, added to the quality of the alignments and to overall translation quality.

2 Domain Specificity

2.1 Word-Association list

Moore (2001) describes a method for learning translation relationship between words from bilingual corpora. The five step process is restated here:

1. Extract word lemmas from the Logical Form created by parsing the raw training data.
2. Compute association scores for individual lemmas.
3. Hypothesize occurrences of compounds in the training data, replacing lemmas constituting hypothesized occurrences of a compound with a single token representing the compound.
4. Recompute association scores for compounds and remaining individual individual lemmas.
5. Recompute association scores, taking into account only co-occurrences such that there is no equally strong or stronger association for either item in the aligned logical-form pair.

The word-association list (WA) was created by applying this method to our training data set of 200,000 aligned French-English sentences of computer manual and help file data. A French linguist determined the best cutoff for the raw data, i.e. determined the association score which would determine the cutoff, and otherwise left the file unedited for inclusion in the transfer training stage. For internal reasons, we used

only associations which are conceptually single word to single word, where a single word can be defined as an item returned as one unit by the analyzer, even though it might be a multi-word item in the source text, e.g *base_de_donnée* <-> *database*. The files included 30,000 pairs, which in their totality, were judged to be 60% accurate[1].

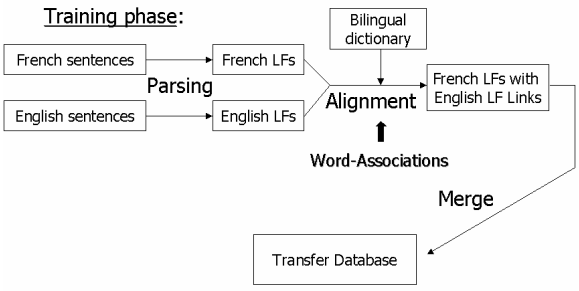

Figure 2

The word association file was used only in training (see Figure 2) to enhance the opportunity for alignment during the detection of transfer patterns.

Examples of WA pairings :

- cliquer click
- processeur CPU
- éclairage lighting
- http://www.mcafee.com http://www.mcafee.com
- nettoyer scavenge
- conversion translation
- Requête/édition query/edit

2.2 Title Association list

The second file used was a specialized file created using the same algorithm, but allowing multi-word titles that are all in capitals in English to associate with multiple words in French that have mixed capitalization on major content words. Because these phrases are identified by using capitalization, they are also referred to as Captoids (Moore, 2001). Items such as *Organizational Units*, which occur with complete capitalization in English, are associated with the French translation, *Unités d' organisation*, a unit which is less easily identified on its own, due to the mixed case.

The information yields approximately 2600 pairs of this type:

Unités d' organisation <->
Organizational Units
Voir aussi <-> *Related Topics*

This title association file (TA) is used in training of the transfer patterns but are also added to the processing of the French training text; they are treated as multi-word lexical entries similar to any French dictionary entry. They become part of the translation dictionary as well. The inclusion of *Voir aussi* as a lexical noun phrase at the analysis stage (French) allows it to parse correctly, and permits the correct translation. Many of the occurrences of Title association pairs are menu names which are syntactically verb phrases (*Voir aussi*) and would have parsed less well without the TA file.

(1)
***Source*:** Pour plus d'informations sur l'utilisation du Gestionnaire de périphériques, consultez *Voir aussi*.
Reference: For more information about using Device Manager, see *Related Topics*.
ALL translation: For more information about using of the manager of devices, see **Related Topics**.
NONE translation: See for more information on using of the Device Manager; *also See.*

However, the evaluation shows that the overall effect of title associations is much less than that of word associations, presumably because the frequency of these items is low in the overall test set.

3 Experiment and Methodology

In order to evaluate the relative quality of the translations with and without the word association and title association strategies, we performed several evaluations of machine translation quality. These evaluations were performed by an independent organization that provides support for NL application development; the evaluators are completely independent of development activities.

We performed two separate sets of evaluations. In the first, we evaluated the full

[1] The size of of the WA file of 42,486 reported in Moore 2001 includes multiple word associations which were not used in this experiment.

version of our system with the Word Association and Title Association components against versions of the system from which we had removed those components. We thus expected that versions of the system with the WA and TA components would outperform those without.

In the second evaluation, we tested the versions of our system with and without the WA and TA components against a benchmark system (the latest release of the French-English Systran system, run with settings appropriate for the computer domain) to see whether the addition of the combination of these components would significantly improve our scores with respect to that benchmark.

3.1 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 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.

4 Results

4.1 Results with multiple versions of our system

In order to isolate the effects of the WA and TA components on the system as a whole, we built 3 new versions of the system:

- *NONE*: Includes neither TA nor WA.
- *No TA*: Includes WA but not TA.
- *No WA*: Includes TA but not WA.

We evaluated each of these versions of the system against our baseline system (ALL), which contains both the WA and TA components. Our hypothesis was that the removal of each of the two components would cause the experimental systems to significantly underperform the ALL system.

We evaluated 250 sentences[2] in each condition in which the output strings for System 1 (ALL) and System 2 (NONE, NoWA, and NoTA, respectively) were not identical. In other words, this analysis shows the amount of improvement between the systems in only those sentences which show any change at all in each condition. For each condition, we calculated the statistical significance of the hypothesis that ALL system is better than the comparison system (e.g. that the score is greater than 0), taking into account both variations in the sentence sample, and variations across the judgments of individual raters.

[2] The data used for testing is blind, i.e. withheld from development and not included in the training set.

Conditio n	Score	Sample Size	Significance
ALL/NONE	0.233 +/- .095	250	> .99999
ALL/NoWA	0.267 +/- .09	250	> .99999
ALL/NoTA	0.063 +/- .093	250	.91

Table 1: Results with differences only

The results show that, for sentences affected by the combination of the WA and TA components, the ALL condition is significantly better than the NONE condition, at a significance level of 0.95. In addition, for sentences affected by the presence of the WA component only, the ALL condition is significantly better than the No WA condition. However, the ALL condition is not significantly better the NoTA condition.

Another question of interest is the effect of the experimental components on the corpus as a whole, rather than just on the sentences that changed; it is possible that the effects we found might have become diluted below the significance threshold because of sparsity of the differences across the whole corpus. Rather than do additional evaluations, we determined the proportion of differences in each condition, and extrapolated a larger sample, assuming that sentences which were absolutely identical would receive a score of 0, using the same 250 judgments as in the previous analysis.

Condition	iffs hecked	otal diffs n test set f 2965	ercent f diffs n test set	Projected sample size to get 250 diffs
NONE /ALL	250	1307	19.13	567
NoWA/ALL	250	1170	21.37	634
NoTA/ALL	250	280	89.29	2647

Table 2: Projected sample sizes

As expected, the results using the projected sample were still positive, though the scores were lower due to the larger sample size. Again, the improvements in the NONE/ALL and NoWA/ALL conditions are significant across the whole data set.

Condition	Score	Sample size	Significance
NONE /ALL	0.103 +/- .04	567	> .99999
NoWA/ALL	0.105 +/- .035	634	> .99999
NoTA/ALL	0.006 +/- .008	2647	.90

Table 3: Results across whole sample

4.2 Results against benchmark system

In a second analysis, we tested to see if the experimental changes to the system improved the performance of our system against our regular benchmark. We selected a random sample of 250 sentences, and translated them using first the ALL, and then the NONE, versions of our system. We also translated them using the benchmark system. We predicted that sentences translated using the ALL system would be significantly better than the sentences translated using the NONE system in its performance against the benchmark.

Condition	Score	Sample size
NONE /benchmark	-0.18 +/- .1	250
ALL/benchmark	-0.14 +/- .11	250

Table 4: Results against Benchmark system.

The difference between these two scores is on the border of significance using a one-tailed paired t-test ($p = .051825$; $t = -1.6334$).

5 Discussion

The premise of the experiment described here was that pairs of translations which were automatically derived from the training data would increase the number of transfer pairings found and improve the quality of translation. The results show that the combination of the word association list and title association list does in fact give us an improvement in quality of translation.

We have measured the change in size in the transfer database, and found that the database shows increased numbers of transfer patterns retained (transfer patterns seen only once were discarded) when the word association file is used, for instance:

Condition	Unique transfers kept
NONE	316518
ALL	368853

Table 5: Increase in patterns kept

We have found from informal observation that increased number of transfers in the transfer database correlates with better performance, particularly if the translation correspondence includes more than one word.

Whereas the WA and TA files have been judged elsewhere on the quality of the translation pairs themselves (Moore 2001), we are primarily interested in whether the data interacts in a positive way with a full-scale automatic alignment process. The result might appear disappointing at first glance, since it is barely significant. However, our experience is that a gain of .04 against the benchmark represents a noticeable difference in quality translation from the user's perspective.

It is important to note as well that this result was achieved even in the presence of a sizeable translation dictionary. We found that the combination of the bilingual dictionary and the structural mapping in the alignment process had already enabled a number of "domain specific" translation correspondences, e.g. *journal* <-> *log* as in example (2) below. In a sense, the alignment algorithm had been able to overcome some domain specific lexical gaps on its own.

The evaluation results give us a number of illustrations of improved transfer patterns. The only difference between the output categorized as NONE and the output categorized as ALL is the use of a transfer database trained with both the WA and TA files included.

(2)

***Source*:** Le tableau ci-dessous explique la fonction des différentes options disponibles dans l'onglet **Journal des transactions** de la boîte de dialogue Propriétés de la base de données.

Reference: This table shows the options and their functions available on the **Transaction Log** tab of the Database Properties dialog box.

ALL translation: The table explains the function of different options available in the **Transaction Log** tab of the dialog Properties box of the database below.

NONE translation: The table explains the function of different options available in the tab **transactions Log** of the dialog Properties box of the database below.

The pattern which caused the improvement is the correspondence (*Journal des transactions* <-> *Transaction Log*) was learned on different pairs of sentences during the alignment phase due to the presence of the word *log* introduced by the word association file.

Without the addition of *log* at alignment time, the alignment process mapped *Journal* to *Log*, but not the more complex mapping for *Journal des transactions*. Compare the translations from the FE dictionary to the pairs from the word association file (where ordering represents frequency of each translation). Note that the WA list has learned the most relevant technical translation (*log*), which was lacking in the FE dictionary, but also the most frequent general translation (*journal*):

FE dictionary
(journal)=(journal magazine diary newspaper)
Word association list
(journal)=(**log** journal newspaper)

A similar case below (3) shows that the inclusion of the word *push* as a translation of *émission* in the word association file allows for a correct pattern in the transfer database:

réplication par émission
push replication

(3)

Source: Pour configurer un serveur WINS afin d'utiliser une **réplication par émission**, vous pouvez faire votre choix parmi plusieurs options configurables de la console WINS.

Reference: To configure a WINS server to use **push replication**, you can choose from several WINS console configurable options.

ALL translation: For configuring a WIN server to use a **push replication**, you can do your

choice among options configurable of the WIN console.

NONE translation: For configuring a WIN server to use a **replication by program**, you can do your choice among options configurable of the console WIN.

FE dictionary: (émission)=(program transmission broadcasting emission broadcast issue uttering)

Word association list: (émission)=(issue **push** Transmit transmit issuance)

This example is quite interesting, because the link of *push* to *émission* is helpful, even though is would be judged incorrect in a standard evaluation of the pairings themselves

We have described the improvements so far as increases in domain specificity, but the effect is more wide-spread. We find that the added information allows for creation and retention of such generally better patterns as those in example (4):

(4)

Source: Assurez-vous qu'il y a du papier dans l'imprimante.

Reference: ***Make sure there is*** *paper in your printing device.*

ALL translation: ***Make sure*** *that* ***there is*** *a paper in the printer.*

NONE translation: ***Provide*** *that a paper in the printer* ***becomes****.*

We note the improved transfer patterns for Make sure and there is.

The incidence of faulty translation patterns learned because of incorrect word-associations has been difficult to measure, but appears to be low. One instance was the learned correspondence of *éteindre* <-> *off* (instead of *turn_off*). We believe this could be avoided by more accurate preservation of information from our Logical Form representation in step one of the Moore algorithm.

5.1 Future improvements

The experiment presented here is the first step in our search for techniques that contribute to the quality of the translations by providing domain specific additions.

We are working to find the most productive method for pruning low accuracy pairs (but still without hand-editing). We have already seen that if the data is truncated to maximize the accuracy of the word associations, the impact on the translation quality drops off, presumably because the high frequency pairs in the word association file contribute fewer unknown translations than the larger noisier file. This suggests that in the process of seeding an automatic alignment process such as ours, recall is more important than precision.

.

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.

Frederking, Robert, et al. 1994. Integrating Translations From Multiple Sources Within the Pangloss Mark III Machine Translation System. In Technology Partnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association of Machine Translation in the Americas. 73-80.

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.

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.