Modular MT with a learned bilingual dictionary: rapid deployment of a new language pair

Jessie Pinkham Martine Smets

Microsoft Research
One Microsoft Way
Redmond, WA 98052
{jessiep martines}@microsoft.com

Abstract

The MT system described in this paper combines hand-built analysis and generation components with automatically learned example-based transfer patterns. Up to now, the transfer component used a traditional bilingual dictionary to seed the transfer pattern learning process and to provide fallback translations at runtime. This paper describes an improvement to the system by which the bilingual dictionary used for these purposes is instead learned automatically from aligned bilingual corpora, making the system's transfer knowledge entirely derivable from corpora. We show that this system with a fully automated transfer process performs better than the system with a hand-crafted bilingual dictionary. More importantly, this has enabled us to create in less than one day a new language pair, French-Spanish, which, for a technical domain, surpasses the quality bar of the commercial system chosen for comparison.

1 Introduction

The phrase "MT in a day" is strongly associated with research in statistical MT. In this paper we demonstrate that "MT in a day" is possible with a non-statistical MT system provided that the transfer component is learned from aligned bilingual corpora (bi-texts), and does not rely on any large hand-crafted bilingual resource. We propose instead to use a bilingual dictionary learned only from the same bi-texts. Section 4.2 describes the creation of the new language pair, French-Spanish, and gives evaluation results. Section 4.1 examines the impact of the learned dictionary on our existing French-English system.

2 Previous work

Commercial systems and other large-scale systems have traditionally relied heavily on the knowledge encoded in their bilingual dictionaries. Gerber & Yang (1997) clearly state that Systran's translation capabilities are dependent on "large, carefully encoded, high-quality dictionaries". With the advent of bi-texts, efforts to derive bilingual lexicons have led to substantial research (Melamed 1996, Moore 2001 for discussion), including resources for semi-automatic creation of bilingual lexica such as SABLE (Melamed 1997), used for instance in Palmer et al. (1998). Statistical MT systems have relied on bi-texts to automatically create word-alignments; in many statistical MT systems however, the authors state that use of a conventional bilingual dictionary enhances the performance of the system (Al-Onaizan et al. 1999, Koehn & Knight 2001). We find then, that in spite of the movement to create bilingual dictionaries automatically, there is still a heavy reliance on hand-crafted and hand-edited resources. We found no full-scale MT system that relied only on learned bilingual dictionaries and certainly none that was found better in performance for doing so.

Rapid deployment of a new language pair has been one of the strong features of statistical MT systems. For example, "MT in a day" was a stated goal of the workshop on statistical MT (Al-Onaizan et al. 1999). The system deployed was of low quality, in part because of the small size of the corpus used, and the difficulty of the language pair chosen (Chinese to English). We have chosen French-Spanish, because we are constrained by the availability of well-developed analysis and generation components in our experiment. Those, needless to say, were

not created in one day, nor were the large size monolingual dictionaries that they rely on. But given the assumption that these modules are available and of good quality, we demonstrate that training the transfer dictionary[1] and example base on bi-texts is sufficient to create a new language pair which is of comparable quality to others based on the same source language. This, to our knowledge, has not been done before in the context of a large hybrid MT system

3 System overview

The MT system discussed here uses a source language broad coverage analyzer, a large multi-purpose source language dictionary, an application-independent natural language generation component which can access a full monolingual dictionary for the target language, and a transfer component. The transfer component, described in detail in Menezes (2001), consists of high-quality transfer patterns automatically acquired from sentence-aligned bilingual corpora.

The innovation of this work is the use of an unedited, automatically created dictionary which contains translation pairs and parts of speech, without any use of a broad domain, general purpose hand-crafted dictionary resource. The architecture of the MT system as described elsewhere (Richardson et al. 2001) used both a traditional bilingual dictionary and an automatically derived word-association file at training time, but it used only the traditional bilingual dictionary at runtime. We refer to this below as the HanC system, because it uses a Hand-crafted Dictionary[2]. We changed this so that a learned dictionary consisting of word-associations (Moore 2001) with parts of speech and a function word only bilingual dictionary (prepositions, conjunctions and pronouns) replaces the previous combination both at training and at runtime[3]. We refer to this as the LeaD system (Learned Dictionary). We demonstrate that this change improves sentences that differ between both systems, and show that we can now adapt quickly to new language pairs with excellent results.

Analysis of the consequences of removing the standard hand-crafted bilingual dictionary from the system (and having no dictionary as a fallback at all) are provided in Pinkham & Smets (2002). It proved important to have a dictionary containing parts of speech to use as a fallback, motivating the work described here.

4 Experiments

We conducted two experiments. In the first one, we compared the performance of the HanC (Hand-Crafted dictionary) MT system to the performance of our LeaD (Learned Dictionary) system. The French-English system is trained on 200,000 sentences in the computer domain, and tested on unseen sentences from the same domain.

In the second experiment, we created a new language pair, French-Spanish, in less than 8 hours. The French-Spanish system was trained on 220,000 sentences from the same computer domain, and also tested on unseen computer domain data.

4.1 French-English translation with a learned bilingual dictionary

4.1.1 Comparing HanC to LeaD

In this first experiment, we compare the performance of the HanC system and the LeaD system for French-English versus the same competitor.

Translations produced by the two versions of our system differ in 30% of the cases. Out of the 2000 sentences in our test set, only 595 were translated differently. In about half of these cases, there was an overt difference in the word chosen as a fallback translation at runtime. In the other half, the translation example-base patterns were different.

[1] In both French-English and French-Spanish, we use a hand-crafted bilingual function word dictionary of about 500 entries. It includes conjunctions, prepositions and pronouns; see section 4.1.4.

[2] The dictionaries are automatically converted from electronic dictionaries acquired from publishers, and are updated by hand over time.

[3] The same statistical techniques identify certain multi-word terms for parsing and transfer. This learned dictionary stays constant during the French-English experiments.

We evaluated 400 of the 595 "diff" sentences mentioned. A complete description of the evaluation method is given in Richardson (2001), and repeated in Appendix A. Evaluation for each version of the system was conducted against the competitor system, which we use as a benchmark of quality. Our current benchmark for French-English is Systran[4], which uses relevant dictionaries available but has not been otherwise customized to the domain in any way.

	Scores	Signif.	Size
HanC system (diffs only)	-.1777 +/-.087	> .999	400
LeaD system (diffs only)	-.0735 +/-.182	.97	400
French-English HanC system	+.2626 +/- .103	> .999	400
French-English LeaD system	+.2804 +/-.115	> .999	400

Table 1: LeaD vs. HanC for FE

We also evaluated a set of 400 sentences taken randomly from the 2000 test sentence set. They were translated with both the HanC system and the LeaD system, and evaluated against the same competitor, Systran.

4.1.2 Results

The random test has a score representative of the quality of the system (December 2001 system), and is significantly better than the competitor given the score of +0.2804 (0 means the systems are the same, -1 the competitor is better, 1 the competitor is worse). See Table 1.

Sentences whose translations differ between the HanC and LeaD versions of our system are less well translated overall. Through examination of the data, we have found that reliance on the fallback translation at runtime tends to indicate a failure to learn or apply transfer patterns from the example-base, both of which are often due to faulty analysis of the source sentence. There are also cases where translations are not learned because of sparse data, but these tend to be rare in our technical corpus.

More importantly, we see that the LeaD version of the system has a significantly higher score than the HanC version (p=0.002 in a one-tailed t-test). Replacing the conventional bilingual dictionary with the learned bilingual dictionary combined with the small function word dictionary has led to significant improvement in quality when measured on "diff" sentences, i.e. cases where all the sentences are different. However, when we take 400 random sentences, the difference between the two versions only affects 30% of the sentences (133 or thereabouts) and therefore does not result in a significant difference (p=0.13 in a one tailed t-test).

4.1.3 Translation examples

In this section, we give examples of translation with both versions of our system, and compared to Systran. The LeaD version of our system uses the correct translation of "casiers", in this specific context, while both our HanC version of the system and Systran use terms inappropriate for this domain. By using a learned dictionary, the LeaD system is better suited to the domain.

Source	Le finisseur est traité comme trois casiers individuels,
Reference	The Finisher is addressed as three individual **bins**
LeaD	The finisher is processed like three individual **bins**.
HanC	The finisher is processed like three individual **pigeonholes**.
Systran	The finisher is treated like three individual **racks**,

4.1.4 Creation of the learned bilingual dictionary

The learned dictionary with parts of speech was created by the same method (Moore, 2001) as the previously used word-association file, with the exception that parts of speech were appended to lemmas in the first step of the process. We are easily able to modify the input this way, because we use the output of the analysis of the training data to create the file that is the input to the word alignment process.

Appending the part of speech disambiguates homographs such as "use", causing them to be

[4] Systran was chosen on the basis of its ranking as the best FE system in the IDC report (Flanagan & McClure, 2000)

treated as separate entities in the word-association process:

use^^Verb
use^^Noun

The word-association process assigns scores to each pair of words. We have established a threshold below which the pairs are discarded. Here are the top word pairs in the learned dictionary for this domain:

utiliser^^Verb use^^Verb
fichier^^Noun file^^Noun
serveur^^Noun server^^Noun

Because the input to the learning process is derived from Logical Forms (the output of our analysis systems), and because this format no longer includes lemmas for function words, there are no function words in the learned dictionaries. This is the primary reason why we complemented the learned dictionary with a function word dictionary. See the future work section for ideas on learning the function words as well.

Both the French-English and the French-Spanish were arbitrarily cut off at the same threshold, and were not edited in any way, resulting in a file with 24,000 translation pairs for French-English and 28,000 translation pairs for French-Spanish. The dictionary for function words contains about 500 word pairs. The traditional French-English dictionary had approximately 40,000 entries.

4.2 French-Spanish

4.2.1 Creating French-Spanish

Our group currently has both a French-English system and an English-Spanish system. In choosing the new language pair to develop, we were constrained by the availability of good quality analysis and generation systems. This is a limiting factor, but will become less so once we have more generation modules available for use[5], as we currently have seven fully developed analysis modules. We were fortunate to have 220,000 aligned sentences for French-Spanish from the technical domain (manuals, help files), which enabled the construction of the learned bilingual dictionaries, and the automatic creation of the transfer pattern example base.

For reasons explained above, our first learned dictionary made no attempt to learn function word translations. We needed, therefore, to complement the learned French-Spanish dictionary with a French-Spanish function word bilingual dictionary, which was bootstrapped from our French-English and English-Spanish bilingual dictionaries. All the translations for prepositions, conjunctions and pronouns were created using both of these, and hand-edited by a lexicographer bilingual in French and Spanish.

The creation process, including the hand-editing work, took less than 8 hours.

4.2.2 Results

The test was conducted on 250 test sentences from the same technical domain as the training corpus, using the methodology described in Appendix A. All test data is distinct from training data and unseen by developers. The Sail Labs French-Spanish system is the benchmark used as comparison. The technical domain dictionary on the website was applied to the Sail Labs translation, but it was not otherwise customized to the domain.

The Sail Labs translation included brackets around unfound words, which were thought to interfere with the raters' ability to compare the sentences; the brackets were removed for the evaluation.

Condition	Scores	Signif	Size
FS LeaD	+.2278 +/- .117	> .999	250
French-English	+.2804 +/- .114	> .999	400

Table 2: French Spanish results

As seen in Table 2, where the French-Spanish system is ranked at +0.228, it is significantly better than the Sail Labs French-Spanish system in this technical domain. The score is very similar to the French-English score as measured against Systran (+.2804). Since these are being compared against different

[5] Members of our group (Corston-Oliver et al.) are developing an automatic generation component. This could speed up the development of generation modules, giving us a potential of 42 different language-pairs trainable on bi-texts.

competitors, we also wanted to measure their absolute quality. On a scale of 1 to 4, where 4 is the best, we found that both Systran and Sail Labs were comparable in quality, and that our system scored slightly higher in both cases, but not significantly so, if one considers the confidence measures (Table 3). The details of the scoring for absolute evaluations are given in Appendix B. As a brief illustration, the LeaD French-English translation in 4.1.3 has a score of 3, while the LeaD French-Spanish translation in 4.2.3 received a score of 2.5.

	Absolute score	
FS LeaD	2.676 +/- .329	250
FS Sail Labs	2.444 +/- .339	250
French-English	2.321 +/- .21	400
FE Systran	2.259 +/- .291	250

Table 3: Absolute scores FS and FE

4.2.3 Translation Example for French-Spanish

This section gives examples of translation from French into Spanish. The LeaD translation has the correct translation for domain specific terms such as "hardware" and "casilla de verificación", while Sails Labs translation does not in spite of the use of a domain bilingual dictionary.

Source	Si la case à cocher Supprimer de ce profil matériel est activée, le périphérique est supprimé du profil matériel.
Reference	Si la casilla de verificación Quitar este perfil de hardware está activada, se ha quitado el dispositivo del perfil de hardware.
LeaD	Si se activa la casilla de verificación Eliminar de este perfil de hardware, el dispositivo se quita del perfil de hardware.
Sails Labs	Si la coloca a marcar Suprimir de este perfil material es activada, el periférico se suprime del perfil material.

5 Future Work

We are planning to experiment with lowering the threshold for the cutoff of information in the learned bilingual dictionary, in an attempt to include more word pairs (some words remain untranslated).

To further validate the Learned Dictionary approach, we are experimenting with other domains. One might assume, for instance, that as the domain becomes broader, learned dictionaries would be less effective due to sparse data. We have preliminary experiments on Hansard French-English data which indicate that this is not the case.

6 Conclusion

We have demonstrated that we can replace the traditional bilingual dictionary with a combination of a small bilingual function word dictionary and a bilingual dictionary learned from bi-texts. This removes the reliance on acquired or hand-built bilingual dictionaries, which can be expensive and time-consuming to create. One can estimate that for any new domain application, this could save as much as 1-2 person years of customization. This also removes a major obstacle to quick deployment of a new language pair.

We believe that high-quality linguistic analysis is a necessary ingredient for successful MT. In our system, it has enabled automation of the transfer component, both in the learning of the bilingual dictionary and in the creation of example-based patterns.

Appendix A: Relative Evaluation Method

For each version of the system to be tested, seven evaluators were asked to evaluate the same set of blind test sentences. For each sentence, raters were presented with a reference sentence, the original English sentence 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. Raters were also shown two machine translations, one from the system with the component being tested, and one from the comparison system (Systran for French-English, Sails Lab for French-Spanish). Because the order of the two machine translation sentences was randomized on each sentence, evaluators could not determine which

sentence was from which system. 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 judgments of quality. The three-way scale also allowed 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 was similarly simple; each judgment by a rater was represented as 1 (sentence from our system judged better), 0 (neither sentence judged better), or -1 (sentence from Systran or Sails Labs judged better). The score for each version of the system was the mean of the scores of all sentences for all raters. The significance of the scores was calculated in two ways. First, we determined the range around the mean which we could report with 95% confidence (i.e. a confidence interval at .95), taking into account both variations in the sentences and variations across the raters' judgments. In order to determine the effects of each stage of development on the overall quality of the system, we calculated the significance of the difference in the scores across the different versions of the system to determine whether the difference between them was statistically meaningful. We used a one-tailed t-test, since our a priori hypothesis was that the system with more development would show improvement (that is, a statistically meaningful change in quality with respect to the competitor).

Appendix B: Absolute Evaluation Method

At the same time as the relative evaluations are made, all the raters enter scores from 1 to 4 reflecting the absolute quality of the translation, as compared to the reference translation given. The grading is done according to these guidelines:

1 unacceptable:
Absolutely not comprehensible and/or little or no information transferred accurately

2 possibly acceptable:
Possibly comprehensible (given enough context and/or time to work it out); some information transferred accurately

3 acceptable:
Not perfect (stylistically or grammatically odd), but definitely comprehensible, AND with accurate transfer of all important information

4 ideal:
Not necessarily a perfect translation, but grammatically correct, and with all information accurately transferred

References

Al-Onaizan, Y & Curin, J. & Jahr, M. & Knight K. & Lafferty, J. & Melamed, D. & Och, F-J, & Purdy, D. & Smith, N. A. & Yarowsky, D. (1999). *Statistical Machine Translation: Final Report*, Johns Hopkins University 1999 Summer Workshop on Language Engineering, Center for Speech and Language Processing, Baltimore, MD.

Corston-Oliver, S., M. Gamon, E. Ringger, R. Moore. 2002. An overview of Amalgam: A machine-learned generation module. To appear in *Proceedings of the International Natural Language Generation Conference*. New York, USA

Flanagan, M and McClure, S. (2000) Machine Translation Engines: An Evaluation of Output Quality, IDC publication 22722.

Gerber, L. & Yang,J. (1997) Systran MT Dictionary Development in the *Proceedings of the MT Summit V*, San Diego.

Koehn, P. & Knight, K. (2001) Knowledge Sources for Word-Level Translation Models, *Proceedings of the conference on Empirical Methods in Natural Language Processing* (EMNLP)

Melamed, D. (1998). Empirical Methods for MT Lexicon Construction, in L. Gerber and D. Farwell, Eds., *Machine Translation and the Information Soup*, Springer-Verlag.

Melamed, D. (1997). A Scalable Architecture for Bilingual Lexicography, Dept. of

Computer and Information Science Technical Report #MS-CIS-91-01.

Melamed, D. (1996). Automatic Construction of Clean Broad-Coverage Translation Lexicons, *Proceeding of the 2nd Conference of the Association for Machine Translation in the Americas* (AMTA'96), Montreal, Canada.

Menezes, A. & Richardson, S. (2001). A Best-First Alignment Algorithm for Automatic Extraction of Transfer Mappings from Bilingual Corpora. In *Proceedings of the Workshop on Data-Driven Machine Translation*, ACL Conference, June 2001.

Moore, R.C. (2001). Towards a Simple and Accurate Statistical Approach to Learning Translation Relationships Between Words. In *Proceedings of the Workshop on Data-Driven Machine Translation*, ACL Conference, June 2001.

Pinkham, J & Smets, M (2002) Machine Translation without a bilingual dictionary Proceedings of the TMI conference, Kyoto, Japan.

Palmer, M. & Rambow, O. & Nasr, A. (1998). Rapid Prototyping of Domain-Specific Machine Translation Systems, in *Proceedings of the AMTA '98*.

Richardson, S. & Dolan, W. & Menezes, A. & Corston-Oliver, M. (2001). Overcoming the Customisation Bottleneck Using Example-Based MT. In *Proceedings of the Workshop on Data-Driven Machine Translation*, ACL Conference, June 2001.