Rapid assembly of a large-scale French-English MT system

Jessie Pinkham[§], Monica Corston-Oliver[°], Martine Smets[§], and Martine Pettenaro[§]

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

[°]Butler Hill Group
http://www.butlerhill.com
moco@butlerhill.com

Abstract

Past research has shown that the ideal MT system should be modular and devoid of language pair specific information in its design. We describe here the assembly of TAMTAM (Traduction Automatique Microsoft), the French-English research MT system under development at Microsoft, which was constructed from a combination of pre-existing rule-based components and automatically created components. At this stage, the system has not been adapted either computationally or linguistically to the French-English context and yet it performs only slightly below the French-English Systran system in independent blind human evaluations

Keywords

Hybrid MT, EBMT, automatic transfer learning, French-English

Introduction

TAMTAM (Traduction Automatique Microsoft) is a French-English translation system which uses a French broad coverage analyzer, a large multi-purpose French dictionary, a 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).

The system is best characterized as a data-driven hybrid system, with rule-based analysis and generation, example-based transfer, and some statistically derived lexical input. The automatic alignment procedure used to create the example base relies on the same parser employed during analysis and also makes use of its own small set of rules for determining permissible alignments. A moderately sized French-English dictionary, containing only word pairs and their parts of speech, provides translation candidates for the alignment procedure and is also used as a backup source of translations during transfer. Statistical techniques supply additional translation pair candidates for alignment and identify certain multi-word terms for parsing and transfer. A complete description of the

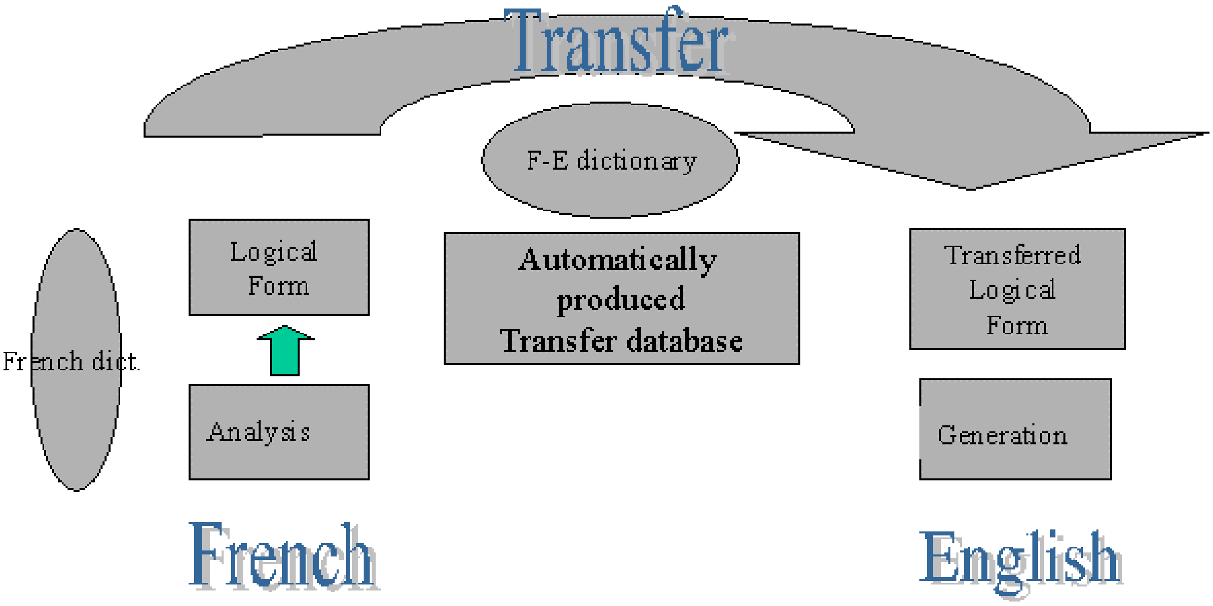


Figure 1

architecture and technical details appears in Richardson et al. (2001).

We will focus here on describing the assembly and early stages of development of the French-English MT system, from its inception in October 2000 to now (April 2001). We report on evaluations conducted so far (in November, February and March), which show significant progress without any hand-coded linguistic changes specific to this language pair.

Review of previous work

Frank (1999) describes an effort to leverage work on parallel grammar development toward Machine Translation. The ParGram project involves parallel development in four languages (English, French, German, Norwegian). The effort stresses the importance of LFG representations, particularly f-structures, which encode functional properties shared across multiple languages, in the construction of an MT prototype covering fewer than 100 sentences. Though we do not rely on the LFG theory as a formalism, we have in our Logical Form component the same advantage of functional/semantic properties shared across several languages.

The transfer rules written for the prototype described by Frank are extremely complex, demonstrating that hand-coded transfer rules would be unsuitable for large-scale MT system development. Frank states that automatic efforts for acquisition of transfer rules and lexical transfer are the direction to explore. TAMTAM takes advantage of parallel development in multiple languages (English and French for this project, but see also Richardson et al. 2001). It also leverages a common functional representation, and incorporates a fully automatic transfer component and bilingual word association learning techniques. Example-based transfer as well as automatically extracted transfer has been proposed in a number of systems (see Somers 1999 for an overview). Richardson et al. (2001) report using fully automated transfer based on bilingual corpora for Spanish-English and English-Spanish that results in commercial level quality. In the current study, using the French-English version of the same system, we focus on studying translation quality at various stages of early assembly, before any language pair specific enhancements.

Modules used to assemble TAMTAM

Assembly of TAMTAM requires the following list of components:

- Aligned bilingual text
- French monolingual dictionary
- Analysis grammar (morphology, syntax, logical form)
- Bilingual FE dictionary
- Automatically derived transfer mappings
- Automatically derived word-association pairs
- English generation component

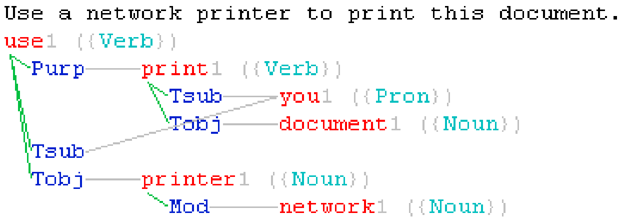
The majority of the components have been described elsewhere (Pinkham 1996, Richardson et al. 2001, Menezes et al. 2001, Moore 2001, Pinkham et al. 2001, Aikawa et al. 2001), but here is a brief outline.

The French dictionary started with a 12,000 word lexicon from Brigham Young University. We manually augmented it to 25,000 entries and added more detailed features on words regarding morphology and syntactic behavior. A Novell word list boosted the number of headwords to about 68,000, and also provided morphological and some syntactic information. The analysis and generation of word forms is enhanced by a morphology engine which comprises both inflectional and derivational components.

The French-English lexicon was built automatically from these sources: Cambridge University Press English-French, Soft-Art 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 47,000 entries and 75,000 translation pairs of words and phrases[1].

We added translation pairs extracted from the actual domain, using statistical word/phrase assignment. The algorithm used is described in Moore (2001). 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 and their interaction with the system are detailed in Pinkham et al. (2001).

The French and English broad coverage parsers produce conventional phrase structure analyses augmented with grammatical relations. Syntactic analyses undergo further processing in order to derive logical forms (LFs), which are graph structures that describe labeled dependencies among content words in the original input. LFs normalize certain syntactic alternations (e.g. active/passive) and resolve both intrasentential anaphora and long-distance dependencies. See Figure 2 for an illustration of a Logical Form representation.



```
Use a network printer to print this document.
use1 ({Verb})
  Purp----print1 ({Verb})
          Tsub----you1 ({Pron})
          Tobj----document1 ({Noun})
  Tsub
  Tobj----printer1 ({Noun})
          Mod-----network1 ({Noun})
```

Figure 2: Sample LF

[1] The work of conversion and inversion of the information was conducted by J. Pentheroudakis.

The code that builds LFs from syntactic analyses is shared between French and English[2]. This shared architecture greatly simplifies the task of aligning LF segments from different languages, since superficially distinct constructions in two languages frequently collapse onto similar or identical LF representations.

TAMTAM acquires transfer mappings by aligning pairs of LFs obtained from parsing sentence pairs in a bilingual corpus. The LF alignment algorithm first establishes tentative lexical correspondences between nodes in the source and target LFs using translation pairs from a bilingual lexicon and the statistical word alignment. After establishing possible correspondences, the algorithm uses a small set of alignment grammar rules to align LF nodes according to both lexical and structural considerations and to create LF transfer mappings. The final step is to filter the mappings based on the frequency of their source and target sides. Menezes & Richardson (2001) provides further details and an evaluation of the LF alignment algorithm.

The target LF produced by the transfer component is mapped to the target string by the rule-based generation component. The generation component has no information about the source language for a given input LF, and the English generation component is thus not specific to TAMTAM.

Our system is very robust: in case syntactic analysis fails, a *fitted parse* is the output of the parser. A fitted parse is a collection of partial parses which cannot be combined together in a single parse by the grammar, but can still be used in further processing. More specifically, it can be used in alignment and transfer, and thus in translation.

In addition to being designed for robustness, most of the components of our system have been tested extensively and are included in commercial products (for example, the grammar checker of Microsoft Word).

Evaluation Method

For each version of the system 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 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 (TAMTAM), and one from the comparison system (Systran[3]). 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 importance of these parameters for subjective 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 TAMTAM judged better), 0 (neither sentence judged better), or -1 (sentence from Systran judged better). The score for each version of the sytstem 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 Systran).

Results

Date of evaluation	Description of system	Score vs. Systran	Sample size
November 2000	Basic components	-.50 +/- .1	308
February 2001	Improved FE dictionary	-.18 +/- .1	250
March 2001	Word-associations	-.14 +/- .11	250

Table 1: Evaluation results

The November system contained only the basic components, while the February system included an improved French-English dictionary. The March system was improved with a dictionary of translation pairs extracted automatically from the domain.

The results of the t-tests show that the February system is significantly better than the November system (t = -.480; p < .00001) at a threshold of .95, while the March to February system comparison is on the border of significance at the .95 level (t = -1.6334; p = .051825). It

[2] LF is shared by all languages under development : English, French, Spanish, German, Chinese, Japanese and Korean.

[3] Systran was chosen as reference system because it is listed in the IDC report (Flanagan, 2000) as the best commercial FE system.

is worth noting, however, that this is in the nature of incremental improvements in system quality; even though each small change may not create a statistically significant improvement, the aggregate of all changes dramatically improves the system over time. We believe that the addition of the word-association components is an example of such a change.

Effectiveness of the transfer mappings

On a set of test data of 500 sentences, using the latest system, TAMTAM used an average of 6.9 mappings per sentence, each spanning approximately 1.6 words. (The greater the span, the more complex the mappings learned.) Table 2 below helps clarify the role played during translation by patterns learned during training

	Transfer database	FE dictionary	Same as source
Lemmas	91.9%	5.3%	2.3%
Pronouns	8.3%	56.1%	
Prepositions	34.5%	65.1%	0.3%
Other relations	38.3%		61.7%

Table 2: Statistics on transfer

TAMTAM derives its transfer information solely from the transfer mappings database and from the FE bilingual dictionary; when both of these fail, the translation is the same as the source[4]. We see that 91.9% of non-function words have translations from transfer mappings, with only 5.3% of cases coming from the FE dictionary. Pronouns are excluded from the alignment-learning algorithm, and yet do appear in the transfer mappings at the rate of 8.3%, when they come in bigger mapping chunks. Prepositions represent key relations which necessarily vary from one language to another, and will be better if they come from the transfer database, where they have been learned from corpora. They are learned 34.5% of the time. Other relations from the LF (Deep Subject, Deep Object) are transferred as is, accounting for the large number of relations that are the same as source (61.7%). We successfully learn other relations 38.3% of the time.

Translation examples

In this section, we present a few examples of translation, and compare them with the translation derived using FE Systran, a commercial translation software which we have taken as benchmark in our evaluations. All settings appropriate for translation in the computer domain were set on in the Systran translations. Our system so far does not outperform Systran on average, although the latest evaluation showed that the systems are quite close in quality of translation. We look at some examples where the raters judged the TAMTAM translation to be better than Systran and give an example of bad TAMTAM translation.

The translation examples are presented as follows: first the source sentence labeled SRC, followed by the reference (human) translation (REF), then the sentence created by our system (TAMTAM), and finally the sentence produced by Systran's machine translation (SYS).

SRC

Dans un réseau basé sur un domaine, les ordinateurs Windows NT sont configurés en tant que membres d'un domaine précis.

REF

In a domain-based network, Windows NT computers are configured as members of a specified domain.

TAMTAM

In a network based on a domain, the Windows NT computers are configured as specific domain members.

SYS

In a network based on a field, the computers Windows NT are configured as members of a precise field.

Even though it uses a domain dictionary, Systran does not select an appropriate translation for the word *domaine* (*domain* in the REF and TAMTAM translations*, field* in the Systran translation).

In addition, the Systran translation fails to produce a English compound NP for the French sequence NN. For example*, les ordinateurs Windows NT* is translated as the *computers Windows NT* by Systran, but as *Windows NT computers* in the human and TAMTAM translations. Note the weakness in the TAMTAM translation: it keeps the definite determiner in English, which is incorrect. This is exactly the type of problem that will be addressed in future linguistic work in the system.

When French uses the NP *de* NP construction, English more often uses a compound NP. TAMTAM uses the automatically learned transfer mappings (no human coding here at all) to correctly generate *specific domain members* for *membres d'un domaine précis*; in the example below, *certification authority* for *Autorité de certification;* and *certificates services* for *services de certificats*. These translations contribute to make correct TAMTAM translation sound more native than the Systran translation, even when both appear to preserve meaning adequately .

SRC

Vous sélectionnez la stratégie qu'une Autorité de certification utilisera lors de l'installation des services de certificats.

REF

You select the policy a CA will use when you install Certificate Services.

TAMTAM

You select the policy that a certification authority will use at the time of installing the certificates services.

SYS

You select the strategy which an Authority of certification will use at the time of the installation of the services of certificates.

But there are worse problems than these in Systran translations. In the example below, *serveurs* is inaccurately translated as *host* by Systran, and this leads to the wrong choice of relative pronoun (*whom*) used here to refer to an inanimate (*server*).

[4] The numbers do not add up to 100% because of another category of untranslatable words, such as numbers, etc.

SRC

Si vous utilisez d'autres serveurs DNS sur votre réseau, vérifiez que l'implémentation de serveur DNS qu'ils utilisent prend en charge les mises à jour dynamiques.

REF

If you are using other DNS servers on your network, verify that they are running a DNS server implementation that supports dynamic updates.

TAMTAM

If you use other DNS servers on your network, verify that the server DNS implementation that they use also supports the dynamic updates.

SYS

If you use other hosts DNS on your network, check that the implementation of host DNS whom they use deals with the dynamic updates.

In the TAMTAM translations, on the other hand, the information automatically collected in the tranfer database and bilingual dictionary permits a correct choice of words and expressions in the target language: *serveur* is translated as *server*.

TAMTAM translations fail mostly when the analysis (parse or LF) is inadequate, sometimes dramatically. In the following example, the clause introduced by Quand is analyzed as a free relative, and the comma after that clause as a coordination. The boundary of the Quand-clause is after *services*, and the main verb does not have a subject. The generation component then inserts a dummy subject*, it*. There is another problem, not related to the bad analysis: *en réduction* is not translated, because it is an unanalyzed unit in French which is not in the bilingual dictionary. On the other hand, some phrases are learned and elegantly translated, such as *le Gestionnaire de services SQL Server* and *le menu Démarrer* (*SQL Server Service Manager* and *Start menu*).

Systran's translation is better but consistently lacks native fluency (and also suffers from untranslated words).

SRC

Quand vous sélectionnez le Gestionnaire de services SQL Server dans le menu Démarrer, l'icône du gestionnaire de services s'affiche en réduction dans la barre des tâches.

REF

When you select SQL Server Service Manager from the Start menu, the Service Manager icon appears minimized in the taskbar by default.

TAMTAM

It appears on the en réduction bar of the tasks when you select the SQL Server Service Manager in the Start menu , the icon of the Manager of services.

SYS

When you select the Manager of services SQL Server in the Démarrer menu, the icon of the Manager of services is displayed in reduction in the bar of the tasks.

We expect that with the constant feedback provided by working on MT, we will be able to address a large number of these errors and improve the analysis modules. We are optimistic that we will then achieve consistently accurate, native-sounding translation.

Conclusion

The resulting French-English system performs slightly less well than Systran, which is the best French-English commercial system according to the IDC report (Flanagan 2000). By studying the TAMTAM system at its inception, we hope to demonstrate the modularity of the Microsoft Research MT system architecture and the remarkably good results achieved with automatic transfer learning alone. We expect a jump in quality once we begin to address language-pair specific linguistic issues in detail.

References

Aikawa, T. & Melero, M. & Schwartz, L. & Wu, A. (2001). Multilingual Sentence Generation. In Proceedings of the workshop on Natural Language Generation, ACL Conference, June 2001.

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

Frank, A. (1999). From Parallel Grammar Development towards Machine Translation – A Project Overview. In Proceedings of the MT Summit VII.

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. (1996). Grammar Sharing in French and English. In Proceedings of the First International Conference on Industrial Applications of Natural Language Processing (IANLP).

Pinkham, J. & Corston-Oliver, M. (2001). Adding Domain Specificity to an MT system. In Proceedings of the Workshop on Data-Driven Machine Translation, ACL Conference, June 2001.

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.

Somers, H. (1999). Review Article: Example-Based Machine Translation. Machine Translation 14: 113-157.

Way, A. (1999). A hybrid architecture for robust MT using LFG-DOP. Journal of Experimental and Theoretical Artificial Intelligence 11(1999)441-471.