Tuning General Translation Knowledge to a Sublanguage

Michael Carl* **and Philippe Langlais**
Laboratoire de Recherche Appliquée en Linguistique Informatique
Université de Montréal, Montréal, Québec, Canada
email:{carl,felipe}@iro.umontreal.ca

Abstract

We present and evaluate an algorithm which extracts a translation grammar (TG) from a bilingually aligned text. We show that lexical transfer rules automatically extracted from specific-domain alignments (an excerpt from the Canadian Hansard) enhance the quality and coverage of a general-purpose bilingual dictionary. We discuss a number of desirable properties of TGs. We show that TGs can be fruitfully integrated within a statistical machine translation system.

1 Introduction

In the framework of example-based machine translation, a number of methods have been proposed for automatically inducing translation correspondences from aligned texts. In so-called "pure" EBMT systems, the only available knowledge resource is the aligned text itself, (Block, 2000; Brown, 2003), while in richer systems additional, linguistic or lexical knowledge resources are used to a varying degree. Experiments have also been carried out where alignments are parsed with the aim of linking word translations and internal nodes of derivation trees in both language sides of the alignment. Wu (1995), for instance, proposes a bilingual stochastic parser. This parser analyses both language sides of an alignment in parallel where the leaves of the binary derivation trees are terminal symbols and all internal nodes are non-terminals. Watanabe (2003; Menezes and Richardson (2003) and Meyers et al. (1998) parse each language side independently and try to find the most probable node correspondences in the derivation trees.[1]

This paper describes and evaluates an algorithm which generates and filters a translation grammar (TG) from partially parsed and aligned texts. Section 2 describes the algorithm and section 3 reports a number of experiments and findings.

The aim of the algorithm is to induce a set of the most probable invertible, compositional and homomorphic transfer rules from an aligned text. We briefly clarify these concepts.

Homomorphy and **Isomorphy**
As defined in (Huijsen, 1997), a TG is homomorphic if every rule has the same arity on its left-hand side (LHS) and right-hand side (RHS) and there exists a 1-to-1 link between the variables in LHS and the variables in RHS. The TG in figure 4, for instance, is homomorphic because every transfer rule has the same number of variables ($NP^{1,2}$ or $DP^{1,2}$) on its LHS and RHS which are linked by their superscribed indexes. Homomorphic TGs generate isomorphic derivation trees, where for every partial derivation tree in one language there exists a linked

*now affilated to: IAI, Saarbrücken, Germany

[1] For an overview of different techniques and representations see (Somers, 2003).

derivation tree with the same depth and arity in the other language.

Compositionality and **non-monotonicity** As Turcato and Popowich (2003) point out, translations can be compositional or they can be non-monotonic. For instance, the translation equivalence (a) is compositional because "viaje" is a translation of "trip" and "negocios" is a translation of "business" while the translation equivalence (b) is non-compositional ("field" is not a translation of "estudio") and (c) non-monotonic combining both, compositional and non-compositional parts.

a)	business trip	↔	viaje de negocios
b)	field trip	↔	viaje de estudio
c)	long field trip	↔	viaje de estudio largo

It is hard for an uninformed learner to know up to what extent a translation is compositional and when it starts to become non-monotonic. Therefore, for every highest scored isomorphic translation, we extract both the most compositional and a non-monotonic transfer rules.

Ambiguity and **Invertibility** In order to achieve high reliability of the TG, every transfer rule in a TG should be unique. As we shall show in a later section, ambiguities produced during transfer which cannot be resolved in the target language do not help to increase the translation quality. For this reason some researchers (Menezes, 2002) retain sufficient context in transfer rules to distinguish it from competing mappings during translation. In this paper we compare invertible TGs and ambiguous TGs. A TG is invertible iff both language sides of the transfer rules are unique in the grammar; ambiguous grammars allow more than one mapping of the same string into the target language.

2 Inducing a TG

This section describes an algorithm which generates a TG from a set of alignments. TGs are made up of lexical transfer rules which contain only terminal symbols and translation templates - i.e. generalized transfer rules - which also contain variables. The resources required for the induction include:

1. a set of reference alignments (RA) from which a TG is to be induced.
2. a (partial) parser for both languages which brackets those translation units to be extracted from the alignments.
3. a bilingual dictionary to find connecting anchor points in the two language sides of the reference alignments.

The algorithm works in three phases:

1. Finding possible chunk-chunk (c-c) translations
2. Generating Translation Templates
3. Filtering the TG

These phases are discussed more closely in the following subsections. The algorithm is depicted in figure 5.

2.1 Finding possible c-c Translations

In the first part of the algorithm in figure 5, each alignment is successively treated to find potential c-c translations. First, bilingual anchors are detected by means of an English-French dictionary. Then a partial parser recognizes complex (nominal) chunks independently in both language sides. The result of this processing is a lexically anchored and partially parsed alignment, as shown in Figure 1.

Potential c-c translation candidates are detected in the alignment (those that are actually present in the general dictionary) and weighted. The weight is the mean of an internal weight w_i and an external weight w_e, the computation of which is described hereafter. Intuitively, the internal weight captures the strength (or ambiguity) of the lexical anchors which a c-c translation contains, while the external weight captures the strength (or ambiguity) of the c-c translation in which it is contained.

The **internal weight** is recursively calculated based on the internal weights of the lexical and the complex anchors it contains (see equation 2 in figure 5). Complex anchors are c-c translations which are contained in larger c-c translations. For instance in Figure 2, NP_{3-4} ↔

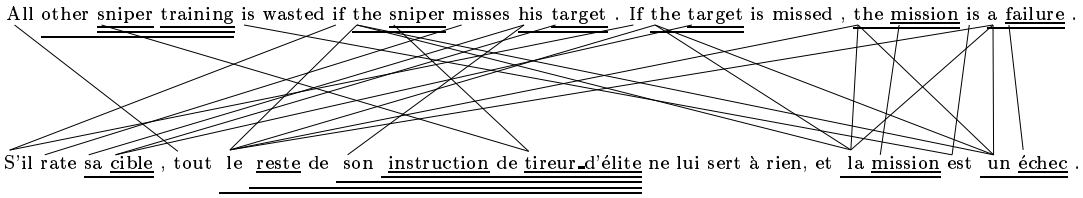

Figure 1: A lexically anchored and partially parsed English-French alignment. Connecting lines represent lexical anchors; underlined words represent chunks.

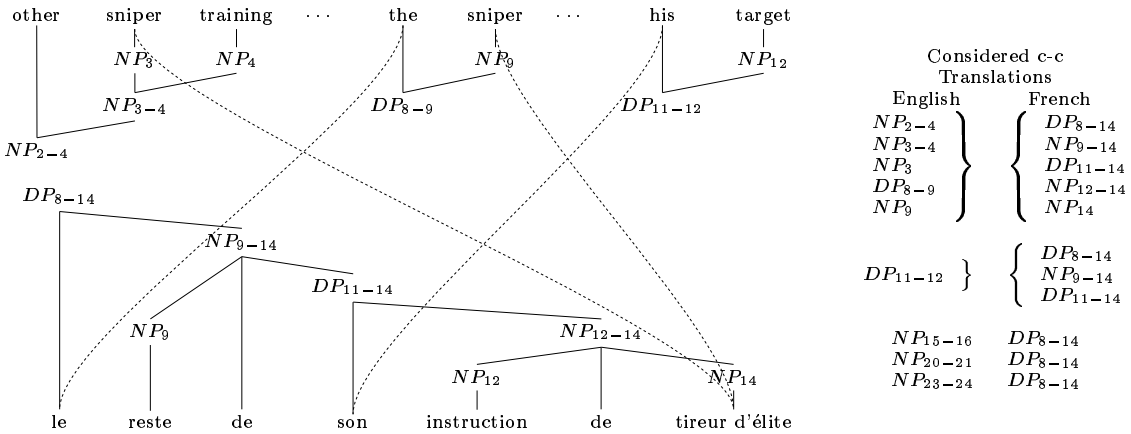

Figure 2: A segment of the alignment in Figure 1 showing three English chunks lexically anchored with one French chunk. Subscribed indexes in nodes refer to the position in alignment of Figure 1.

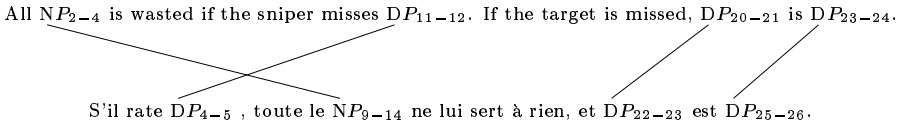

Figure 3: Template of the Alignment with substituted Chunk-Chunk (c-c) Translations

1 All other sniper training is wasted if the sniper misses his target. If the target is missed, the mission is a failure.
↔
S'il rate sa cible toute le reste de son instruction de tireur d'élite ne lui sert à rien, et la mission est un échec.

2 All DP^1 is wasted if the sniper misses DP^2. If the target is missed, DP^3 is DP^4.
↔ S'il rate DP^2 toute le NP_1 ne lui sert à rien, et DP^3 est DP^4.

3 (other sniper training)$_{NP}$ ↔ (reste de son instruction de tireur d'élite)$_{NP}$

4 (other NP^1 NP^2)$_{NP}$ ↔ (reste de son NP^2 de NP^1)$_{NP}$

5 (training)$_{NP}$ ↔ (instruction)$_{NP}$

6 (sniper)$_{NP}$ ↔ (tireur d'élite)$_{NP}$

7 (his target)$_{DP}$ ↔ (sa cible)$_{DP}$

8 (his NP^1)$_{DP}$ ↔ (sa NP^1)$_{DP}$

9 (the mission)$_{DP}$ ↔ (la mission)$_{DP}$

10 (the NP^1)$_{DP}$ ↔ (la NP^1)$_{DP}$

11 (a failure)$_{DP}$ ↔ (un échec)$_{DP}$

12 (a NP^1)$_{DP}$ ↔ (un NP^1)$_{DP}$

Figure 4: Generated Translation Grammar for the English-French alignment of Figure 1. Superscribed indexes in nodes indicate links of variables.

$$s(lhs \leftrightarrow rhs) = 2 * \frac{w(lhs \leftrightarrow rhs)}{\sum w(lhs \leftrightarrow \_) + \sum w(\_ \leftrightarrow rhs)} \quad (1)$$

$$w_i(lhs \leftrightarrow rhs) = \begin{cases} 2 * \frac{\#(lhs \leftrightarrow rhs)}{\#lhs + \#rhs}, & \text{if } lhs \leftrightarrow rhs \text{ lexical anchor} \\ \frac{anch(lhs,rhs)}{1+noise(lhs,rhs)}, & \text{if } lhs \leftrightarrow rhs \text{ complex anchor} \end{cases} \quad (2)$$

$$w_e(lhs \leftrightarrow rhs) = \sum \frac{w_i(Lhs, Rhs)}{\#D(Lhs \leftrightarrow Rhs)} \quad if \begin{cases} lhs \text{ is substring of } Lhs \\ \text{AND} \\ rhs \text{ is substring of } Rhs \end{cases} \quad (3)$$

$$anch(Lhs, Rhs) = \sum w_i(lhs \leftrightarrow rhs) \quad if \begin{cases} lhs \text{ is substring of } Lhs \\ \text{AND} \\ rhs \text{ is substring of } Rhs \end{cases} \quad (4)$$

$$noise(Lhs, Rhs) = \sum w_i(lhs \leftrightarrow rhs) \quad if \begin{cases} lhs \text{ is substring of } Lhs \\ \text{XOR} \\ rhs \text{ is substring of } Rhs \end{cases} \quad (5)$$

Finding possible chunk-chunk (c-c) translations:
1 for all bilingual alignments $LHS \leftrightarrow RHS$:
2 match alignment against bilingual dictionary to find lexical anchors.
3 calculate internal weights w_i of lexical anchors according to equation 2.
4 parse nominal expressions (i.e. chunks) independently in LHS and RHS.
// an example of the resulting representation is shown in figures 1 and 2
5 for all anchored c-c translations $lhs \leftrightarrow rhs$ in $LHS \leftrightarrow RHS$:
6 begin
7 recursively compute an internal weight $w_i(lhs \leftrightarrow rhs)$ based on the internal weights of the c-c translation(s) contained in $lhs \leftrightarrow rhs$ according to equation 2.
8 recursively compute an external weight $w_e(lhs \leftrightarrow rhs)$ based on the internal weights of the c-c translations in which $lhs \leftrightarrow rhs$ is contained, according to equation 3
9 set $w(lhs \leftrightarrow rhs) += (w_i(lhs \leftrightarrow rhs) + w_e(lhs \leftrightarrow rhs))/2$
10 end
11 end

Generation of Translation Templates:
12 for all anchored c-c translations $lhs \leftrightarrow rhs$: compute c-c score $s(lhs \leftrightarrow rhs)$ according to equation 1
13 for all anchored c-c translations $lhs \leftrightarrow rhs$:
14 generate a set of translation templates $tls \leftrightarrow trs$ by substituting the highest ranked c-c translations which are contained in $lhs \leftrightarrow rhs$
15 set $w(tls \leftrightarrow trs) += (w_i(tls \leftrightarrow trs) + w_e(tls \leftrightarrow trs))/2$
16 end

Filter a TG:
17 for all translation templates $tls \leftrightarrow trs$: compute scores $s(tls \leftrightarrow trs)$ according to equation 1
18 for all bilingual alignments $LHS \leftrightarrow RHS$:
19 find highest ranked and consistent translation template $TLS \leftrightarrow TRS$ which is a generalization of $LHS \leftrightarrow RHS$
20 recursively print most compositional template $TLS \leftrightarrow TRS$ and most specific transfer rule $LHS \leftrightarrow RHS$.
21 end

Figure 5: Algorithm for inducing a Translation Grammar from bilingual alignments. #D($Lhs \leftrightarrow Rhs$) stands for the number of daughters of the rule.

NP_{12-14} (the c-c translation "sniper training $\leftrightarrow$ instruction de tireur d'élite") is a complex anchor in $NP_{2-4} \leftrightarrow NP_{9-14}$ because it is anchored by the lexicon entry "sniper $\leftrightarrow$ tireur d'élite" and NP_{3-4} and NP_{12-14} are substrings in NP_{2-4} and NP_{9-14} respectively.

The internal weight of lexical anchors is the Jaccard's coefficient (Reijsbergen, 1979) shown in equation 2 (upper). For instance, the bilingual dictionary contains the translation "sniper $\leftrightarrow$ tireur d'élite" which is assigned an internal weight of 1.0 since "sniper" and "tireur d'élite" occur only once in the dictionary. The translation "the $\leftrightarrow$ le" has a weight of 2/3 because "the" can translate into "le" or "un" but "le" only translates into "the". Note that anchors are computed based on the string equality of the lemma(s).

The internal weight of a c-c translation takes into consideration its anchors (lexical and complex) and the noise produced by those anchors which connect to different chunks. For instance, the English-French c-c translation $NP_{3-4} \leftrightarrow NP_{12-14}$ in Figure 2 has a lexical anchor "sniper $\leftrightarrow$ tireur de élite" and a complex anchor $NP_3 \leftrightarrow NP_{14}$. It has the 'noisy anchors' $NP_9 \leftrightarrow NP_{14}$ and $NP_{8-9} \leftrightarrow NP_{14}$.

In order to re-distribute weights into c-c translations an **external weight** w_e is calculated for each c-c translation which is contained in a larger anchored chunk. For instance, the translation "training $\leftrightarrow$ instruction" in Figure 2 inherits proportional weights from all anchored c-c translations which dominate "training" and "instruction". The external weight is thus a way to distribute — recursively from the sentence level — weight mass over word translations, also those which are not covered by the general dictionary, therefore acting as a smoother.

2.2 Generation of Translation Templates

In the second part of the algorithm, generalizations are induced from the detected c-c translations. First scores are calculated for each c-c translation as shown in line 12 of the algorithm in Figure 5. Then a limited number of translation templates is generated by substituting the most probable complex anchors in c-c translations. Since non-overlapping complex anchors are simultaneously substituted in LHS and RHS of the cc-translations, both sides of the resulting template have the same number of linked variables. In this way homomorphy for the translation grammar is generated.

2.3 Filtering a TG

First the scores of the translation templates are computed (line 17 in Figure 5). Similar to section 2.1, the loop runs over the set of alignments and not, as in section 2.2, over the set of potential c-c translations. For each alignment $LHS \leftrightarrow RHS$, the most likely translation template is retrieved (line 19 in Figure 5) and the most compositional derivations as well as their non-monotonic translations are filtered and added to the TG. In Figure 4, transfer rule 1 is the least general (i.e. non-monotonic) transfer rule that can be extracted from the alignment while transfer rule 2 is the most compositional rule which is consistent with the aligned text and the extracted TG.

The TG in Figure 4 contains 12 transfer rules which have been generated and filtered from the alignment in Figure 1. Note that a variant, the transfer rule 3a (below), was not extracted as this would have implied the translation template 4a. Translation templates 4a and 4b are ambiguous since their left-hand sides are identical but their right-hand sides are not. As template 4b has a higher score than template 4a in the reference alignments, the latter one was suppressed in favor of the former, and template 4 was filtered.

3a	sniper training	$\leftrightarrow$	instruction de tireur d'élite
4a	other NP^1	$\leftrightarrow$	reste de son NP^1
4b	other NP^1	$\leftrightarrow$	autre NP^1

Even though the translation "training $\leftrightarrow$ instruction" was not contained in the initial bilingual dictionary, this entry is given quite a high score due to its frequency and the explanatory strength it has to support isomorphic translation. In the same way, the lexical anchor "sniper $\leftrightarrow$ tireur de élite" has an internal weight of 1, because it occurs only once in the dictionary. However, due to the variations in which the French

and English forms appear in the text, the filtered translation has only a score of 0.65.

3 Experiments

This section reports a number of experiments which are based on the algorithm described in section 2. First we present an experiment to induce a large-scale TG and discuss the resources involved. Then we use the lexical transfer rules of the induced grammar to translate a test text from English into French. Further, we show that the induction algorithm is scalable to different sizes of reference alignments. We compare the translation quality produced when using ambiguous and invertible TGs, we compare it with a general purpose MT-system (BabelFish) and a standard statistical machine translation (SMT) engine. Finally we combine the SMT system with the TGs. We conclude by summarizing the major findings of the experiments.

3.1 Extraction of a TG

In this experiment, we gauge the potential of generating a TG from large domain-specific aligned text. We used 50,000 English $\leftrightarrow$ French reference alignments taken from the Canadian Hansard Corpus. This, we shall say is our set RA_1. The alignments had an average length of 18.8 and 19.9 words for the English and French side respectively. Both texts were PoS tagged and lemmatized while keeping the alignment information intact. The reference alignments contain 13,629 and 13,278 different lemmas for English and French while the French text has 1.3 times more (different) surface forms than the English text (see table 2). Clearly, lemmatization reduces the number of different tokens, makes the texts more comparable and thus better suited for extracting homomorphic translation knowledge.

To detect lexical anchors in the alignments, we used a bilingual English $\leftrightarrow$ French dictionary containing 77,016 entries with 49,341 and 45,695 different lemmas for the English and French side respectively. Most of the entries are unique while some words are highly ambiguous: the French word "support", for instance has 57 entries while the English word "support" has 34.

The dictionary covered almost 3/4 of the words in RA_1 using only 7,688 and 7,714 different lemmas for English and French respectively. That is, only 15.6% and 16.9% of the lemmas contained in the dictionary also occurred in RA_1. In addition, 42.3% and 43.5% of the English and French words in the alignments were anchored in the bilingual dictionary. That is, more than 20% of the words in RA_1 were only in one language side of the dictionary but without a lexical connection into the other language side.

We used KURD (Carl et al., 2002) as a partial parser, taking as input the PoS tagged and lemmatized text. A set of 15 and 17 rules detected simple and complex nominal expressions up to a recursion depth of 6 for the two languages. 581,599 French chunks and 650,136 English ones have been generated in RA_1.

The algorithm described in section 2 had as an output the TG_1 containing more then 3.6 transfer rules on average for each alignment: 113,810 lexical transfer rules and 70,153 translation templates (see table 2).

3.2 Translating a Test Text (TT)

In this experiment we evaluate the TG by translating (from English into French) a test text (TT) of 500 sentences taken from the Canadian Hansard Corpus, but from a different period than the alignments from which we have generated the grammar[2]. We compare over a translation reference the coverage and the accuracy of two translations; one produced by only looking up a general-purpose dictionary (the same lexicon used for anchoring during subsentential alignment), and the other one obtained by looking up the automatically induced lexical transfer rule.

Comparing the translations produced using TG_1 and the bilingual dictionary (DIC), no significant progress can be observed with respect to the coverage of the source text. Both the dictionary and TG_1 cover around 67% of the English words, as shown in table 1.

[2]The average sentence length of the test alignments is similar to that of RA_1.

Table 1: Coverage of TG_1 and a Dictionary (DIC)

	TG_1	DIC
#words	8,665	9,806
#covered words	5,752	5,796
%covered words	66.38	66.99
BLEU	0.1421	0.0573
WER	68.89%	81.68%
SER	93.2%	99.6%
chunks len ≥ 2		
#chunks	966	146
#covered words	2,652	325
%covered words	30.61	3.75

A different picture emerges if we examine the translation quality in Table 1, in terms of fully automatic computed scores, BLEU (Papineni et al., 2002), WER (for Word Error Rate, computed as a classical edit-distance) and SER (for Sentence Error Rate, which is the ratio of translations that are not verbatim the gold-standard translation). Clearly an increase in translation quality can be observed for TG_1. More interestingly, the coverage of chunks with length ≥ 2 increases from 3.75% as for the general purpose dictionary, to more than 30% in TG_1, which suggests a correlation between the length of the matching chunks and the translation quality.

3.3 Scalability of Grammar Induction

In this experiment we wanted to know whether the induction algorithm is scalable to different number of reference alignments.

We extracted a set RA_0 containing 10,000 alignments and a set RA_2 containing 100,000 alignments from the Canadian Hansards. The maximum length of these alignments was limited to 30 words, while in RA_1 no such limitation existed and the longest alignment had more than 80 words[3]. In all three sets of reference alignments, the ratio of words per lemma is around 1.30 for English and between 1.64 and 1.78 words/lemma for French. The number of words in the three sets and their distribution is shown in table 2. The table compares also the three generated translation grammars $TG_{0,1,2}$.

[3]The length of an alignment is the maximum number of words in *LHS* or *RHS*.

Table 2: Scalability to different size of reference alignments

English $\leftrightarrow$ French Reference Alignments

	RA_0	RA_2	RA_1
#alignments	10,000	100,000	50,000
#words in E	151,954	1,437,450	938,078
#words in F	163,113	1,503,196	997,194
#diff. words E	7,343	22,501	17,915
#diff. words F	9,528	29,559	23,675
#diff. lemma E	5,663	17,260	13,629
#diff. lemma F	5,796	16,736	13,278

Invertible Lexical Transfer Rules

	TG_0	TG_2	TG_1
#transfer rules	23,214	180,745	113,810
#rules used for TT	3,581	4,685	4,405
#covered words in TT	4,611	6,146	5,752

Quality of Translated Test Text (TT)

	TG_0	TG_2	TG_1
WER	71.91%	66.93%	68.89%
BLEU	0.1365	0.1704	0.1421

The average number of extracted lexical transfer rules decreases slightly as the number of reference alignments increases, from 2.3 rules per alignment in RA_0 to 1.8 rules per alignment for RA_2. The average length of the extracted transfer rules increases from 8.8 and 9.5 words/entry in TG_0 to 10.3 and 11.1 words/entriy in TG_2 for English and French respectively.

According to the scores BLEU and WER, higher quality translations is obtained as more reference translations are available for generating the grammar. Also the coverage of the source text increases from 53% (4,611 words) for TG_0 to 71% (6,146 words) for TG_2 with 3,581 and 4,685 chunks matched respectively. While around 50% of the translated English source words are covered by chunks of length 1 the number of words covered by chunks of length ≥ 2 climbs from 22% for TG_0 to 33% for TG_2. Here, again, we see a correlation between length of matching chunk and translation quality.

3.4 Comparing Invertible/Ambiguous TGs

In this experiment we examine the impact of ambiguous TG on the translation quality. While the translation grammars $TG_{0,1,2}$ in the previous experiments were invertible in the sense defined in section 2, in this experiment we induce ambiguous translation grammars $TG^a_{0,1,2}$ from the three sets of reference alignments $RA_{0,1,2}$. To generate ambiguous TGs, we set a threshold in lines 19 and 20 of the algorithm in Figure 5 such that ambiguous transfer rules are extracted if their score is higher than or equal to 10% of the most probable and consistent translation template. In this way we allow ambiguous transfer rules as shown in Table 3.

Table 3: Ambiguous Transfer Rules

the commission	↔	la commission
the commission	↔	le conseil
the commission	↔	une commission
last year	↔	an dernier
last year	↔	année dernière

As expected, the number of entries in $TG^a_{0,1,2}$ increases compared to the invertible $TG_{0,1,2}$; and this by around 12%. Since almost all different words (and lemmas) of the reference alignments were already contained in the invertible TGs the ambiguous versions do not contain significantly more different tokens. Nevertheless, the coverage of the test text increases, specially considering chunks of length ≥ 2.

Table 4: Ambiguous Translation Grammars

	TG^a_0	TG^a_2	TG^a_1
#transfer rules	28,393	220,248	146,684
WER	71.88%	67.22%	69.75%
BLEU	0.1398	0.1706	0.1519

However, comparing the results in table 4 and 2, there is no clear indication as to whether the translation quality produced by the ambiguous TGs is better than the quality of the invertible TGs. For instance, TG^a_1 obtains a better (i.e. higher) BLEU score than TG_1, but a lower WER value. We thus conclude that, adding ambiguities into a TG does not lead to better translation quality — at least not if the ambiguities cannot be resolved during translation.

3.5 Comparison with SMT and BabelFish

This experiment is designed to compare the translation quality of the TGs with the quality of a statistical MT system (SMT) and with Systran's BabelFish (BF).

Table 5: Comparing TG with SMT and BF

System	BLEU	WER	SER
BF	0.1578	66.03%	97.87%
SMT_0	0.1156	74.72%	95.04%
SMT_1	0.1231	73.54%	96.69%
SMT_2	0.1378	71.52%	94.80%
SMT_3	0.2061	61.66%	93.38%

We trained the SMT system[4] with the tree reference alignments $RA_{0,1,2}$ discussed in section 3.1 and 3.3. In addition, we used a 15 times bigger reference set RA_3, containing 1,6 million alignments from the Canadian Hansards. We then translated the test text (TT) based on these four models. The outcome is shown in Table 5. A clear increase in quality can be observed as the number of reference alignments increases. However, a direct comparison of the Tables 5 and 2, shows that $TG_{0,1,2}$ alone get better BLEU scores and a lower WERs than the SMT system when trained on the same reference alignments. This result was to be expected, since many more resources are used for the induction of $TG_{0,1,2}$ than for estimating a standard IBM translation model.

On the other hand, we were surprised by the observation that Systran's Babelfish yields scores inferior to SMT_3 and even inferior to TG_2. One reason for this seems to be due to the fact that Babelfish is a general purpose translation system which lacks the typical translation knowledge contained in and extracted from the Canadian Hansards. Translations such as "the speaker/le président" or "some hon. members:

[4] a noisy channel engine relying on IBM2 model, a trigram language model and a DP decoder, see (Langlais and Simard, 2002) for more details.

oh, oh !/des voix: oh, oh" are typical for the Canadian Hansards and could be produced by the trained systems but have not be produced by BF. Despite this, we had the impression that BF translation were of better quality than either of the other systems. This, on the other hand, indicates that automatic evaluation metrics might been appropriate for learning systems but not necessarily so for general purpose rule-based MT systems.[5]

3.6 Integrating TG and SMT

In this experiment we wanted to see whether SMT and TG could be integrated and whether an integration would add to the translation quality. The integration of the TG and SMT was based on the approach described in (Langlais and Simard, 2002): the SMT is constrained to use the translation contained in the TG, if a sequence in the source language string matches the entry in the TG. For instance, if the TG contains the lexical transfer rule "the commission $\leftrightarrow$ la commission" and a sequence in the English source sentence matches the LHS expression then the French translation generated by the SMT system has to contain the target expression "la commission".

As can be seen in Table 6, with this integration, the translation quality improves in every case where the same set of reference alignments are used in SMT and TG. From this we conclude that both techniques cover complementary properties which can be fruitfully integrated for translation.

Table 6: Integrating TG and SMT

	BLEU	WER	SER
SMT_0-TG_0	0.1495	71.19%	93.61%
SMT_1-TG_1	0.1684	70.32%	91.73%
SMT_2-TG_2	0.1789	68.94%	92.20%
SMT_3-TG_1	0.1928	67.99%	90.93%
SMT_3-$TG_1 \geq 2$	0.2096	64.51%	90.93%

However, when integrating the much larger statistical translation model SMT_3 with the lexical transfer rules of TG_1, we observe a decrease in quality. We therefore investigated a second integration version: forcing SMT_3 to use translation proposals from TG_1 only if their length ≥ 2, but whether or not this yields a better translation quality is unclear (better BLEU score, but worse WER).

The reason why TG_1 does not enhance the translation quality when linked with SMT_3 seems to to be that most of the translations for the collocations captured in chunks of length ≥ 2 were also produced by the SMT_3 system. For instance both the SMT_3 and the TG_1 were producing "le chef de l'opposition" as a translation of "the leader of the opposition". Therefore the TG_1 did not provide additional knowledge to the SMT_3 engine. However, this must be balanced by the fact that the SMT_3 system was trained on a training corpus 30 times bigger than TG_1.

Finally, we wanted to see whether the integration of ambiguous translation proposals from the $TG^a_{0,1,2}$ would still enhance the translation quality. From the ambiguous proposals, the SMT_3 system would have to choose one translation which it considers best suited in the context of the generated target language string. Provided with the choice of generating one of "la commission", le conseil" or "une commission" as in Table 3, and given that the SMT_3 system's French target language model was trained on a huge text, the idea was to let it work as a disambiguation module. The translation quality (as measured by BLEU and WER) was however lower than the one produced by the invertible grammars.

4 Conclusions

We have presented a "knowledge-rich" algorithm which allows to generate a translation grammar (TG) from an aligned text. We have shown that a TG extracted from different corpora (10,000 to 100,000 alignments) exceeds and refines by far the translation knowledge contained in a general purpose dictionary.

We have shown that the induction algorithm is scalable to different reference alignment sizes and that an ambiguous TG does not have an

[5] The validation of this hypothesis is far beyond the scope of the present study.

advantage over an invertible one, if no mechanism is provided to resolve ambiguities at runtime. We also observed that our grammar induction procedure produces higher quality translations than a "knowledge-poor" approach (i.e. a standard statistical machine translation system) when both systems are fed with the same amount of reference alignments.

To our surprise, we have also observed that a general purpose MT-system (BabelFish) lags behind the measured translation quality of the trained systems. We found two main reasons for that: the domain-specific jargon which BabelFish did not have the chance to adapt to, and the automatic metrics that we used to evaluate the quality of the translation.

Last but not least, we have also shown that the combination of both approaches yields better translations. We conclude that the methodology offers a way to acquire domain specific translation knowledge which could be fruitfully applied in an MT system.

References

Hans Ulrich Block. 2000. Example-Based Incremental Synchronous Interpretation. In *(Wahlster, 2000)*, pages 411–417.

Ralf Brown. 2003. Transfer rule Induction for Example-Based machine translation. In *(Carl and Way, 2003)*.

Michael Carl and Andy Way, editors. 2003. *Recent Advances in Example-Based Machine Translation*. Kluwer, Academic Publisher, Boston/Dordrecht/London. in press.

Michael Carl, Johann Haller, Christoph Horschmann, Dieter Maas, and Jörg Schütz. 2002. The TETRIS Terminology Tool. *TAL, Structuration de terminologie*, 43(1).

Willem-Olaf Huijsen. 1997. Translation completeness for context-free grammars. In *WEBSLS, The European Student Journal of Language and Speech*. http://www.essex.ac.uk/websls/papers/97-04/97-04.html.

Philippe Langlais and Michel Simard. 2002. Merging example-based and statistical machine translation: An experiment. In *Proceedings of the fifth Conference of Association for Machine Translation in the Americas (AMTA)*, pages 104–114, Tiburon, California, oct.

Arul Menezes and Stephen D. Richardson. 2003. A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In *(Carl and Way, 2003)*.

Arul Menezes. 2002. Better contextual translation using machine learning. In *Machine Translation from Research to Real Users*, pages 124–134, Berlin Heidelberg. Springer.

Adam Meyers, Roman Yangarber, Ralph Grishman, Catherine Macleod, and Antonio Moreno-Sandoval. 1998. Deriving transfer rules from dominance-preserving alignments. In *Computerm, First Workshop on Computational Terminology*, Montreal, Canada.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 311–318, Philadelphia, Pennsylvania, USA, jul.

Cornelis van Reijsbergen. 1979. *Information Retrieval*. Butterworths, London. http://www.dcs.gla.ac.uk/Keith/Preface.html.

Harold Somers. 2003. An Overview of EBMT. In *(Carl and Way, 2003)*.

Davide Turcato and Fred Popowich. 2003. What is Example-Based Machine Translation. In *(Carl and Way, 2003)*.

Wolfgang Wahlster, editor. 2000. *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer, Heidelberg.

Hideo Watanabe. 2003. Finding translation patterns from dependency structures. In *(Carl and Way, 2003)*.

Dekai Wu. 1995. Grammarless extraction of phrasal translation examples from parallel texts. In *TMI-95*.