Using Source-Language Transformations to Address Register Mismatches in SMT

Manny Rayner, Pierrette Bouillon
University of Geneva, FTI/TIM
40 bvd du Pont-d'Arve, CH-1211 Geneva 4, Switzerland
{Emmanuel.Rayner,Pierrette.Bouillon}@unige.ch

Barry Haddow
School of Informatics, The Informatics Forum, 10 Crichton Street
Edinburgh EH8 9AB, University of Edinburgh, Scotland
bhaddow@inf.ed.ac.uk

Abstract

Mismatches between training and test data are a ubiquitous problem for real SMT applications. In this paper, we examine a type of mismatch that commonly arises when translating from French and similar languages: available training data is mostly formal register, but test data may well be informal register. We consider methods for defining surface transformations that map common informal language constructions into their formal language counterparts, or vice versa; we then describe two ways to use these mappings, either to create artificial training data or to pre-process source text at run-time. An initial evaluation performed using crowd-sourced comparisons of alternate translations produced by a French-to-English SMT system suggests that both methods can improve performance, with run-time pre-processing being the more effective of the two.

1 Introduction

The most common problem when doing Statistical Machine Translation in the real world is that there isn't enough data, and the second most common problem is that there isn't enough of the right kind of data; in other words, that there is a mismatch between training and test. In this paper, we will look at an example of a common type of mismatch, which arises within the context of the European Framework ACCEPT project. ACCEPT[1] is concerned with the subject, rapidly growing in importance, of translating the content of online user forums. Given the large variety of possible technical topics and the limited supply of online gurus, it frequently happens that users, searching forum posts online, find that the answer they need is in a language they do not know.

Currently available tools, for example Google Translate, are of course a great deal better than nothing, but still leave much to be desired. When one considers that advice given in an online forum may not be easy to follow even for native language speakers, it is unsurprising that a Google Translated version often fails to be useful. There is consequently strong motivation to develop an infrastructure explicitly designed to produce high-quality translations. ACCEPT intends to achieve this by a combination of three technologies: monolingual pre-editing of the source; domain-tuned SMT; and monolingual post-editing of the target. The manual pre- and post-editing stages are performed by the user communities which typically grow up around online forums. In addition, the SMT stage can optionally be bracketed between automatic pre- and post-editing stages.

In this paper, we will only consider the automatic stages of the translation process in the French-to-English translation pair; we wish to translate French forum data for the benefit of English-speaking users. This rapidly exposes a mismatch between training and test data at the level of register. Forum posts are typically informal in tone. The vast majority of available aligned French/English training data is however formal: a typical example, which we will

[1] Automated Community Content Editing PorTal; http://www.accept.unige.ch.

use in the rest of the paper as our primary resource, is the proceedings of the European parliament, the ubiquitous Europarl corpus (Koehn, 2005). Similar problems would have arisen if we had used other corpora, e.g. the UN corpus[2], Callison-Burch's giga corpus[3] or the Canadian Hansard corpus[4].

French is a language where the gap between formal and informal usage is large. (For purposes of comparison, it is much larger than in English, though perhaps not as large as in Arabic). We will focus on two immediate problems, verb forms and questions. French, like most European languages (English is the major exception) has two second-person pronouns, the formal *vous* and the informal *tu* (accusative form *te*, elided to *t'* before a vowel). Each pronoun has multiple different associated verb inflections. Thus for example the present, future and subjunctive forms of *voir* (to see) are *voyez*, *verrez* and *voyiez* for the formal pronoun, but *vois*, *verras* and *voies* for the informal one. Question-formation is also linked to the distinction between formal and informal register. In the informal register, questions are often formed using the expression *est-ce que*, e.g. *Est-ce que vous avez des pommes?*, "Do you have apples?". In the formal register, the question is usually formed by inversion of subject and verb, so here *Avez-vous des pommes?* If the subject is not a pronoun, this requires introduction of a dummy subject, e.g. *Votre ami a-t-il des pommes?*, "Does your friend have apples?", literally "Your friend does he have apples?"

In some contexts, for example literary translation, it would be important to maintain register differences when translating French to English, perhaps translating *Est-ce que tu veux venir?* as "You wanna come?" but *Voulez-vous venir?* as "Do you want to come?" In the context of forum chat, where the central issue is obtaining useful help, this seems an overrefinement. In what follows, we will assume that we can freely translate informal French constructions as formal English constructions, which will make it much easier to reuse formal-register training data.

Table 1 presents figures, contrasting numbers of occurrence of *tu*, *te* and *est-ce que* in the French-English Europarl version 6 corpus (formal register) and ACCEPT forum logs (informal register). As can be seen, *tu*, *te* and *est-ce que* are all common words in ACCEPT, but much less common in Europarl. The limited quantity of training data for informal-register constructions results in problematic translations when they occur in ACCEPT test data; for example, the question-formation phrase *est-ce que* is often translated literally as "is it that", and informal second-person verbs often turn out to be out-of-vocabulary.

	Europarl		ACCEPT	
tu	273	0.015%	5421	6.9%
te	105	0.006%	2154	2.7%
est-ce que	1109	0.061%	212	0.3%
vous	90564	4.9%	4022	5.1%
questions	62031	3.4%	6588	8.4%
#sents	1825077		77819	

Table 1: Numbers of sentences containing occurrences of *tu*, *te*, *est-ce que* and *vous*, as well as numbers of questions, in the French-English Europarl corpus (formal register) and ACCEPT forum logs (informal register).

In the rest of the paper, we will describe experiments in which we have attempted to address these problems by means of source-language rewriting rules which transform informal register constructions to corresponding formal-register ones. The rewriting rules are written in a minimal regular-expression based transduction notation, which only requires access to a good source of lexical information. Transformation rules can be used either at training time or at run-time. At training time, rules can be used to create artificial training data by transforming existing formal-register corpus material into informal-register. Alternately, at run-time, rules working in the opposite direction can be treated as a pre-processing stage, applied before use of the SMT engine, which transforms informal-register phrases into formal-register counterparts. We present the rules themselves and then the results of the experiments.

[2] http://www.euromatrixplus.net/multi-un/

[3] http://www.statmt.org/wmt10/training-giga-fren.tar

[4] http://www.isi.edu/natural-language/download/hansard/

2 Creating rewriting rules

We begin by considering rules for creating artificial training data. At the end of the section, we briefly consider how to invert these rules to construct rules that can be used at run-time.

2.1 Rules for creating artificial data

Our starting point was the French/English Europarl corpus, which contains 1.8M aligned sentence pairs; we began by writing rules which transformed French sentences not containing the lexical items of interest to us (*tu*, *te* and *est-ce que*) into sentences, equivalent in meaning, which did contain these items. Since the transformed sentence has the same meaning as the original one, it can safely be aligned against the same English sentence, to create more training data.

We obtain the lexical information we need from the MMORPH system (Petitpierre and Russell, 1995). The French MMORPH lexicon contains about 2.1M surface forms, each associated with a feature-value list encoding grammatical information. A typical MMORPH verb entry is the following one for *affirmeriez*, the second person plural conditional form of *affirmer*, "to affirm":

```
"affirmeriez" = "affirmer"
Verb[ mode=conditional
tense=present number=plural
person=2 form=surface
type=1 derivation=ant ]
```

For ease of use, the MMORPH lexicon is converted into Prolog form; to increase efficiency, a little preprocessing is carried out to cache some relationships between surface words, in particular the relationship between corresponding second person singular and plural forms of verbs. Transduction is performed by a simple interpreter, also implemented in Prolog; the interpreter will be released as open source and consists of less than two pages of straightforward code. Rules permit matching of regular expressions, where words can optionally be annotated with Prolog calls to access lexical information.

Figure 1 shows a typical rule, which maps a combination of *vous* and a formal second person plural verb to *tu* and an informal second person singular verb, taking account of possibly accompanying words which may be between the subject and the verb, or immediately after the verb. Reading the rule from top to bottom, we have an occurrence of *vous* replaced by *tu*; an optional negation particle; an optional clitic, which if it is *vous* (reflexive object) is replaced by *te*; an optional second clitic; a formal second person plural verb form, which is replaced by the corresponding informal second person plural verb form using information taken from MMORPH; an optional following verb; and an optional *vous-même*, replaced by *toi-même*.

The only non-obvious aspect of the rule is the element

```
(From:2p_verb_indic(From, To))/To
```

which maps the formal second-person plural verb form `From` to the corresponding informal second-person singular form `To`. For most verb forms, the mapping is unambiguous: thus for example *accueillerez* (second person plural future) maps to *accueilleras* (second person singular future). There is however a systematic ambiguity in the ending `-iez`, which can be either the second person plural present or the second person plural subjunctive. The element specifies that, in the case of an ambiguity, the indicative form should be chosen, as opposed to the subjunctive.

The rule is combined with two similar but more complex rules, which match the common contexts requiring a subjunctive verb and try to map the second-person verb to a subjunctive form if possible. The two subjunctive rules are attempted first, and the indicative one from Figure 1 is only used if they fail. Since subjunctive readings are considerably less frequent than indicative ones, and surface cues for identifying subjunctives are fairly reliable, the combination of the three rules performs well in practice; we will justify this claim in the next section.

The design illustrates both the strengths and the weaknessness of the methodology. On the negative side, the rules integrally depend on *ad hoc* properties of French, in this case the relatively unambiguous second person plural verb inflections and the fact that subjunctive contexts can reliably be predicted using a small set of cue words. Given that these conditions are in fact met, the upside is that we are able to produce a simple set of rules which can be efficiently applied. In general, the methology works if

```
[vous/tu,                           %vous -> tu
 opt(or(ne, ’n\’’)),                %optional negation
 opt(or(vous/te, (C:clitic(C))/C)), %optional clitic (vous -> te)
 opt((C1:clitic(C1))/C1),           %optional 2nd clitic
 (From:2p_verb_indic(From, To))/To, %2 pl verb (transformed pl to sg)
 or(V:verb(V, _), []),              %opt verb
 or(’vous-même’/’toi-même’, [])     %opt vous-même -> toi-même
]
```

Examples:

Je pense que cela a été fait parce que **vous n’en avez** pas parlé. →
Je pense que cela a été fait parce que **tu n’en as** pas parlé.
(I think this has been done because you haven’t talked about it)

Vous l’avez dit vous-même, Monsieur le Commissaire.
Tu l’as dit toi-même, Monsieur le Commissaire.
(You have said it yourself, Mr. Commissioner).

Figure 1: Transduction rule for converting second-person plural verbs into corresponding second person singular forms and typical examples of applying the rule, with matched portions in **bold**. Comments are prefaced by percent signs (%). The rule has been slightly simplified for presentational purposes.

it is possible to exploit *ad hoc* properties of this kind to create rules that trigger on reasonably large numbers of sentences that can be transformed into useful variants.

In the experiments described here, we use a total of 12 formal-to-informal rules. In addition to the three described above, we have one rule for creating examples of the accusative informal second-person pronoun *te*, and seven for creating examples of the informal question construction *est-ce que*. The rule for creating examples of *te* is very simple: it matches a sequence consisting of *vous* immediately followed by a verb which is *not* a second-person plural form. This means that the occurrence of *vous* must be an accusative form (*vous* is ambiguous between nominative and accusative, like English *you*), and hence it can here be safely replaced by *te*. The rules for *est-ce que* look for several different versions of sequences, characteristic of questions involving inverted word-order, consisting of a verb followed by a hyphen and a subject pronoun; they rearrange them into corresponding sequences using *est-ce que*, for example mapping *voulez-vous ... ?* (“want-you ... ?”) into *est-ce que vous voulez ... ?* (“*est-ce que* you want ... ?”)

2.2 Run-time rules

The first group of rules above, which transform *vous* to *tu* with changes to the associated verbs, are easy to reverse, and have almost the same form. The reversed versions, which map *tu* to *vous*, can thus be applied at runtime. Since *te* is unambiguous, the reversed rules can safely be extended so that *te* is also mapped to *vous*, in effect creating a set of rules which reverse and combine the first and second groups.

The rules for *est-ce que* are much less complete, only covering certain specific contexts involving pronouns, and are not straightforward to reverse. We will discuss the issues concerned at the end of § 3.4.

3 Experiments

We conducted two groups of experiments, using both the formal-to-informal and the informal-to-formal sets of rules. In the first group, we apply the formal-to-informal rules to the Europarl corpus create artificial training data, and then retrain the ACCEPT SMT models; in the second group, we use the informal-to-formal rules to pre-process ACCEPT data, and then use the baseline SMT models.

In both cases, intuitive assessment of the results suggests that use of the rules often has a positive effect, but the difference in BLEU is small. We consequently performed a contrastive evaluation, presenting pairs of differing translations to judges and asking them to mark which of the two candidate translations they preferred. Judges were recruited through the Amazon Mechanical Turk.

In the rest of this section, we first describe the evaluation methodology (§ 3.1), then the creation of the baseline SMT system (§ 3.2), the group of experiments where the rules were used to create artificial training data, (§ 3.3), and finally the group where the rules were used at run-time (§ 3.4).

3.1 Contrastive evaluation using the Amazon Mechanical Turk

To evaluate the difference in performance between two versions of the French-to-English ACCEPT system on a given corpus, we perform the following analysis. We extract all triples $\langle$source, translation$_1$, translation$_2\rangle$ for which translation$_1$ and translation$_2$ are different. Triples are divided into groups of 20 and posted as HITs on the Amazon Mechanical Turk, offering a payment of $1 per HIT. We limited participation to workers resident in Canada (a country which has both French and English as official languages), requesting only people who were native speakers of English with a good knowledge of written French, and who had moreover already completed at least 50 HITs of which at least 80% had been accepted by the poster. We required three separate judges for each HIT.

Each judge sees the $\langle$source, translation$_1$, translation$_2\rangle$ displayed with translation$_1$ and translation$_2$ presented in a random order, with all the diverging words highlighted in red. Since the average length of a ACCEPT sentence is about 17 words, but the number of highlighted words in a translation is usually between 1 and 4, this greatly simplifies the judge's task. For each triple, the judge chooses between five possible results: first clearly better, first slightly better, about equal, second slightly better, second clearly better. We aggregated results using majority judgements, scoring the result as “unclear” if there was no majority. We evaluate significance of results by applying the McNemar sign test to the aggregated numbers of “better” and “worse” judgements. To estimate inter-judge agreement, we marked groups of judgements as one of the following:

Unanimous All three judgements were identical.

Agree Either no one preferred the first translation or no one preferred the second translation.

Strong disagree One judge strongly preferred the first translation and one strongly preferred the second.

Weak disagree Remaining cases.

The average judging time for a 20-item HIT was 6 minutes and 15 seconds, corresponding to an hourly rate of $9.63, good payment by AMT standards. The strong inter-annotator agreement figures we present below suggests that judges were pleased with the conditions offered and worked conscientiously. Restricting judges to a bilingual country appears to be important. We tried removing this condition, and obtained faster turnaround time but much poorer-quality results, with weak inter-annotator agreement and many anomalous judgements suggesting that judges lacked fluency in one or the other language or were not taking the job seriously.

3.2 Training Data and SMT Systems

The SMT baseline system was a phrase-based system trained with the standard Moses pipeline (Koehn et al., 2007), using GIZA++ (Och and Ney, 2000) for word alignment and SRILM (Stolcke, 2002) for the estimation of 5-gram Kneser-Ney smoothed (Kneser and Ney, 1995) language models.

For training the translation and lexicalised reordering models we used the releases of europarl and news-commentary provided for the WMT12 shared task (Callison-Burch et al., 2012), together with a dataset from the ACCEPT project consisting mainly of technical product manuals and marketing materials. This last data set covers the same topics as the forums we wish to translate (so it may be considered as “in-domain”) but it is almost exclusively in the formal register.

For language modelling we used the target sides of all the parallel data, together with approximately 900,000 words of monolingual English data extracted from web forums of the type that we wish to translate. Separate language models were trained on each of the data sets, then these were linearly interpolated using SRILM to minimise perplexity on a heldout portion of the forum data.

For tuning and testing, we extracted 1000 sentences randomly from a collection of monolingual French forum data (distinct from the monolingual English forum data), translated these using Google Translate, then post-edited to create references. The post-editing was performed by a native English speaker, who is also fluent in French. This 1000 sentence parallel text was then split into two equal halves (`devtest_a` and `devtest_b`) for minimum error rate tuning (MERT) and testing, respectively.

3.3 Creating artificial training data

In our first group of experiments, we applied all the formal-to-informal rules to the French half of the French/English Europarl corpus, creating about 80K transformed pairs; since all the rules transform sentences into paraphrases of themselves, the English side of the pair can be left unchanged. Table 2 summarises the data produced. The transformation process involves three passes, one for each type of rule, and takes a total of about 15 minutes on a high-end laptop.

Type	#Pairs
tu	37184
te	20926
est-ce que	21814

Table 2: Transformed French/English Europarl data produced by rewriting rules of different types.

We added the new artifically produced data to the existing set and retrained the ACCEPT SMT models using the expanded data set. We created two different retrained models: the first (**TU/TE**) added only the *tu* and *te* corpora, and the second (**EST-CE-QUE**) added only the *est-ce que* data. In order to facilitate comparisons with **BASELINE**, we did not re-run MERT for the **TU/TE** and **EST-CE-QUE** systems; we re-used the weights from the **BASELINE** system.

Initial evaluation using BLEU on a held-out set of 500 French forum sentences gave inconclusive results; BLEU was slightly better for **TU/TE** and slightly worse for **EST-CE-QUE**, but the difference was in neither case statistically significant. Since we were only attempting to improve performance on a set of words that occurred in about 10% of the sentences in the corpus, this was unsurprising. In order to concentrate on the phenomena of interest, we randomly extracted a set of 200 ACCEPT sentences containing *tu* or *te*, and 200 containing *est-ce que*, from the monolingual corpus of French forum sentences, distinct from any of the data sets used so far. We will refer to these two test corpora as `tu_200` and `est_ce_que_200`. We processed each corpus using both **BASELINE** and the appropriate version of the retrained system, and subjected the results to comparative judging using the methodology described in § 3.1. The results are summarised in Tables 3 and 4, where in each case we give the figures for aggregated comparisons and inter-annotator agreement, as defined in § 3.1.

Aggregated judgements	
Judgement	Number
BASELINE better	34
TU/TE better	68
Unclear	8
Same result	90
Significance	$p < 0.002$

Inter-annotator agreement	
Agreement	Number
Unanimous	64
Agree	18
Weak disagree	28
Strong disagree	0

Table 3: Comparison between **BASELINE** and **TU/TE** SMT models on `tu_200` test corpus, judged by three AMT-recruited judges.

As can be seen, the two sets behave quite differently. Table 3 shows a solid improvement for **TU/TE** compared to **BASELINE**, with 68 better against 34 worse; however, Table 4 indicates a slight *decline* for **EST-CE-QUE**, 43 to 53. Examination of the data shows that the **TU/TE** is succeeding primarily because it is able to fill lexical gaps, most obviously second-person verb forms that did not appear in **BASELINE**'s training data. **EST-CE-QUE**, in contrast, is able to fill very few lexical holes. The

Aggregated judgements	
Judgement	Number
BASELINE better	53
EST-CE-QUE better	43
Unclear	10
Same result	94
Significance	(not significant)

Inter-annotator agreement	
Agreement	Number
Unanimous	40
Agree	30
Weak disagree	32
Strong disagree	4

Table 4: Comparison between **BASELINE** and **EST-CE-QUE** SMT models on `est_ce_que_200` test corpus, judged by three AMT-recruited judges.

expression *est-ce que* can usually only be translated well when a substantial amount of context is taken into account. The literal translations “is that” or “is it that” are not completely wrong, and attempts to improve on these often just make things worse; adding more training data with examples of *est-ce que* confuses the system as often as it helps it. Thus, although we find positive examples like[5]:

Source: Est-ce que je peux installer NIS2011?
Trans$_1$: Is that I can install NIS2011?
Trans$_2$: Can I install NIS2011?

we get even more negative ones like:

Source: Est-ce que je dois acheter une licence?
Trans$_1$: Is that I have to purchase a license?
Trans$_2$: Can I must purchase a license?

Although it seems to us that there are still interesting possibilities to explore here, the second person singular/plural transformation holds out more immediate promise of concrete gains, and we consequently decided to focus on it in the second set of experiments.

3.4 Applying rules at run-time

Given that the main effect of the artificially created informal second-person data is to fill lexical holes, and that the relevant transformation rules can readily be reversed, it is natural to investigate the idea of using the reversed rules at run-time (cf. § 2.2). This idea is easy to implement: we apply the reversed rules as part of the automatic pre-processing stage (cf. description of the ACCEPT architecture in § 1), replacing *tu* and *te* with *vous* and changing associated second-person singular verbs to the corresponding second-person plural forms in contexts licensed by the rules. The result is then submitted to the **BASELINE** SMT engine. Table 5 shows the result of performing these operations on the `tu_200` test set, evaluated as before by comparing against the plain result obtained without pre-processing. As can be seen, the margin of improvement (87 to 35) is even greater than the 68–34 given by adding artificial training data (Table 3 above).

As a sanity check, we asked the evaluators to compare the results of applying pre-processing directly against the result of adding artificial training data (Table 6) and also applied pre-processing to the `devtest_b` set (cf. § 3.2), comparing it against the plain result for this set (Table 7). Reassuringly, judges confirm that pre-processing is better than adding artificial data (66–34), and that application of pre-processing rules to the general devtest set produces a small improvement (31–10).

Aggregated judgements	
Judgement	Number
Non-pre-processed better	35
Pre-processed better	87
Unclear	8
Same result	70
Significance	$p < 0.0001$

Inter-annotator agreement	
Agreement	Number
Unanimous	77
Agree	23
Weak disagree	20
Strong disagree	10

Table 5: Comparison between plain and pre-processed versions of `tu_200` test corpus, translated by **BASELINE** SMT model and judged by three AMT-recruited judges.

Finally, it is important to note that the non-trivial contexts which most of the rules possess are essen-

[5] In the following two examples, Trans$_1$ is the translation produced by **BASELINE** and Trans$_2$ that produced by **EST-CE-QUE**.

Aggregated judgements	
Judgement	Number
TU/TE/non-pre-processed better	34
BASELINE/pre-processed better	66
Unclear	12
Same result	88
Significance	$p < 0.002$
Inter-annotator agreement	
Agreement	Number
Unanimous	68
Agree	18
Weak disagree	24
Strong disagree	2

Table 6: Comparison between plain version of tu_200 test corpus translated by **TU/TE** SMT model and pre-processed versions translated by **BASELINE** SMT model, judged by three AMT-recruited judges.

tial. In order to test this, we constructed a trivial set of informal-to-formal transformation rules, which simply map every second person singular word (*tu*, *te*, second person singular verb forms, etc) to the corresponding second person plural form. The result was very bad, since, without the constraining contexts, the rules seriously overmatch. Table 8 shows a comparison between the rules used in the main experiments (i.e. with context) and the trivial rules without context.

We had originally intended to carry out similar experiments using reversed versions of the rules for *est-ce que*, but initial investigations convinced us that the problems involved are much more challenging in nature. There are two difficulties. First, the formal-to-informal rules we defined for *est-ce que* only work for examples where the subject is a pronoun, which is the minority case; in the est_ce_que_200 corpus, less than 30% of the examples have a pronominal subject. Second, and even more seriously, the inverted rules would transform *est-ce que* into constructions with an inverted subject, but it is not in fact clear that this transformation improves the quality of SMT translation. Our overall judgement was that a simple approach of the kind we used successfully for tu/vous has almost no chance of succeeding.

Aggregated judgements	
Judgement	Number
Non-pre-processed better	10
Pre-processed better	31
Unclear	4
Same result	466
Significance	$p < 0.002$
Inter-annotator agreement	
Agreement	Number
Unanimous	23
Agree	11
Weak disagree	8
Strong disagree	3

Table 7: Comparison between plain and pre-processed versions of devtest_b corpus, translated by **BASELINE** SMT model and judged by three AMT-recruited judges.

4 Conclusions and further directions

Register mismatches are a common problem in SMT, normally arising because training data is formal register and test data is informal register. We have presented an initial study carried out on a French-to-English SMT system, using source-language rewriting rules both to create artificial training data and as a run-time pre-editing step. We created rules for two common constructions typical of informal-register French language: second-person singular verb forms, and question-formation using *est-ce que*.

Perhaps the most interesting aspect of the study is the very different performance we obtained for the two phenomena. For second-person singular verb forms, both creation of artificial training data and run-time pre-processing worked well, with clear improvements on sentences containing these lexical items; run-time pre-processing appears to be the more effective of the two methods. Our guess is that there are many similar cases, both in this language pair and others, which can be handled using similar methods. The prerequisites seem to be the following:

- Existence of equivalent formal-register words that the informal-register words can be replaced by;

Aggregated judgements	
Judgement	Number
Pre-processing with context better	90
Pre-processing w/o context better	27
Unclear	2
Same result	81
Significance	$p < 0.0001$

Inter-annotator agreement	
Agreement	Number
Unanimous	87
Agree	17
Weak disagree	14
Strong disagree	1

Table 8: Comparison between use of pre-processing rules with and without context on `tu_200` test corpus, translated by **BASELINE** SMT model and judged by three AMT-recruited judges.

- Good SMT translation of the formal-register counterparts; and
- Availability of surface patterns that can identify relevant occurrences of the informal-register words.

In particular, it seems reasonable to us to suppose that the methods would port to other source languages which use different verb-forms for formal and informal language.

The successful treatment of second-person singular verb forms, however, contrasts sharply with the completely unsuccessful attempt to use the same methods on *est-ce que*. We used the rules to create about 20K extra aligned pairs of training sentences. Hand-examination of the artificial data showed that it was of good quality, and yet it not only failed to improve the translation of *est-ce que*, but even degraded it slightly. The problem is the non-local and highly context-dependent translation of *est-ce que*; this depends on the following main verb, which may be widely separated from it. Thus, for example[6], in

Est-ce que la destination de sauvegarde est sur un disque externe? →
Is the save destination on an external drive?

the translation of *est-ce que* becomes "Is" because the following verb is *est*, while in

Est-ce que quelque chose vous semble bizarre avec mon réseau? →
Does something seem strange to you about my network?

the translation is "Does...seem", because the following verb is *semble*.

With enough training data, it is possible that an SMT engine would be able to learn these patterns robustly, but our impression is that a great deal of data would be needed. A more promising approach seems to be to write runtime transduction rules, operating both pre- and post-translation, which perform the necessary regularizations of the source and target language word-orders, as for example described in (Nießen and Ney, 2004). We will be investigating this idea in the near future.

Acknowledgements

The work described in this paper was performed as part of the Seventh Framework Programme ACCEPT project, under grant agreement 288769.

References

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia, editors. 2012. *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Montréal, Canada, June.

R. Kneser and H. Ney. 1995. Improved backing-off for m-gram language modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, volume 1*.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL Demo Session*.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit 5*.

S. Nießen and H. Ney. 2004. Statistical machine translation with scarce resources using morpho-syntactic information. *Computational linguistics*, 30(2):181–204.

[6] The following examples are taken from the `est_ce_que_200` corpus.

F.J. Och and H. Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong.

D. Petitpierre and G. Russell. 1995. Mmorph-the multext morphology program. *Multext deliverable report for the task*, 2(1).

A. Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Seventh International Conference on Spoken Language Processing*. ISCA.