Translating with Scarce Resources

Yaser Al-Onaizan, Ulrich Germann, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Daniel Marcu, and Kenji Yamada
Information Sciences Institute
University of Southern California
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292
{yaser,germann,ulf,knight,koehn,marcu,kyamada}@isi.edu

Abstract

Current corpus-based machine translation techniques do not work very well when given scarce linguistic resources. To examine the gap between human and machine translators, we created an experiment in which human beings were asked to translate an unknown language into English on the sole basis of a very small bilingual text. Participants performed quite well, and debriefings revealed a number of valuable strategies. We discuss these strategies and apply some of them to a statistical translation system.

Introduction

Corpus-based approaches to machine translation (MT) have been on the rise recently, partly because of their promise to automate a great deal of dictionary construction and rule writing, partly because they simply represent a new way of attacking a stubborn problem, and partly because they have performed relatively well in MT evaluations (such as those performed by DARPA and the German Verbmobil program). These approaches generally rely on a large bilingual text corpus to provide sample translations. A statistical model is trained on the samples, and it is used to translate new sentences. Most of this research can be classified as either statistical machine translation (SMT) or example-based machine translation (EBMT); it includes work such as (Brown *et al.* 1993; Nagao 1984; Wu 1997; Alshawi, Buchsbaum, & Xia 1997; Melamed 2000; Och, Tillmann, & Ney 1999).

In this area, it is a truism that "there's no data like more data." If a program sees a particular word or phrase one thousand times during training, it is more likely to learn a correct translation pattern than if it sees it ten times, or once, or never. Corpus-based MT approaches have so far been applied to situations where large amounts of bilingual text already exist. For example, (Brown *et al.* 1993) exploit the substantial French-English Canadian parliamentary record. For many language pairs of interest, such a large corpus does not exist, and this severely limits the practicality of statistical approaches for pairs like Polish/Finnish, German/Romanian, etc. This problem is particularly acute for *low-density* languages, which typically have: few native speakers, low levels of computerization, writing not yet standardized, little linguistic work, and/or few linguistic resources online.

Of course, a small amount of bilingual text can be commissioned outright. Current techniques do not work very well on such small data sets, but this does not rule out the possibility of developing new techniques that do. It is also possible that such new techniques would be applicable to improving performance on large data sets.

In this paper, we make an initial investigation into corpus-based translation with severely limited bilingual corpora. For our experiments, we use *Tetun*, one of fifteen distinct languages spoken in East Timor.[1] While carrying out these experiments, we had no background on the Tetun language. This significant blinder forced us to view the linguistic data from an angle similar to that of a knowledge-poor computer program.

Training Corpus

We were able to obtain a small Tetun/English bilingual corpus from Internet sources. These included a United Nations site with a variety of Timor-related legal documents written in up to four languages (English, Tetun, Bahasa Indonesia, and Portuguese). They also included Australian and Japanese sites specializing in East Timorese studies and humanitarian relief. Our bilingual corpus contained:

- 1102 sentence pairs.
- 23652 English word tokens.
- 25576 Tetun word tokens.
- 1993 distinct English word forms.
- 1729 distinct Tetun word forms.

[1] Another frequently-used spelling is *Tetum*. East Timorese languages break down into many dialects. We focussed on Tetun-Dili, which is spokenly widely in the East Timorese capital.

1. You can find out where you can go to register in several ways.

   Imi sei hatene oin sa mak atu bele ba hetan fatin tau naran hirak ne'e.
2. UNAMET is responsible for running the popular consultation where you - the people of East Timor - will choose the future of East Timor.

   UNAMET sei responsabiliza atu halao konsulta popular nebe imi - povu Timor Loro Sae - sei hili futuru Timor Loro Sae nian.
3. They will also watch over the whole process to make sure the rules are obeyed and the process is fair.

   Sira mos sei tau matan ba prosesus ne'e tomak atu hatene lolos katak ema lao tuir duni ordem no prosesu ne'e justu duni.

Figure 1: Sample sentence pairs from a small Tetun/English bilingual corpus.

We can contrast this with the 1.6 million sentence pairs available in the French/English Hansard corpus.

Because of the small size of the corpus, it was possible to sentence-align it by hand. We could not be sure that the alignment was correct, of course, but we had many clues: corresponding sentence lengths, sentence pairs containing the same proper nouns, matching paragraph boundaries, matching section markings, etc. The fact that we were able to align the Tetun/English corpus is consistent with positive empirical results reported for knowledge-poor sentence-alignment methods (Church 1993). Figure 1 shows a sample.

Testing Corpus

We held out a short news report from the training corpus and divided it into ten sentences. We circulated the Tetun version (shown in Figure 2), but kept the English translation hidden. One question we wanted to ask was:

> *With only a small bilingual training corpus, and no knowledge of the Tetun language (no native speaker, informant, dictionary, grammar book, etc.), can a person translate the material in Figure 2 accurately and fluently?*

If the task is impossible for people, then it is probably asking too much to develop new MT techniques for dealing with small training sets. On the other hand, if the task can be accomplished well by people, then post-task debriefings may shed light on potential strategies for MT.

Another way to assess the impact of corpus size is to examine how familiar the test corpus looks after we have seen the training corpus. In this case:

- 18% of the distinct Tetun word forms in the test data are never observed in the bilingual training data.
- 41% of the distinct Tetun word forms in the test data are observed fewer than six times in the bilingual training data.
- 63% of the distinct Tetun word pairs in the test data are never observed.
- 90% of the distinct Tetun word triples in the test data are never observed.

The figures are higher if we consider tokens rather than distinct word forms. This situation is rather difficult in comparison to a similar example in our French/English corpus, where 99.6% of the test words were observed in training, as were 93.5% of the test word-pairs.

1. Funsionariu senior UNAMET sira ba Maliana, Suai no Viqueque iha Kuarta-feira hamutuk ho Embaixador Agus Tarmidzi xefe Forsa Serbisu (Forsa Tarefa) Indonezia nian, Brigadeiru Jeneral Satoris, Ofisial Polisia (Polri) iha Timor Lorosa'e nian no Ofisial Senior Indonezia nian sira seluk.
2. Sr Ian Martin, Reprezentante Espesial Sekretariu Jeral nian ba Timor Lorosa'e esplika, "Ami ba fatin tolu ne'e tanba fatin hirak ne'e maka fatin sira be ami iha preokupasaun boot liu."
3. Sr Martin hateten katak, "Maske prosesu tau naran nian la'o di'ak, iha akontesimentu seriu boot ida foin lais ne'e iha Viqueque, tanba nune'e ema barak maka halai husik hela sira nia uma.
4. Ida ne'e maka situasaun ida ne'ebe presiza haree."
5. Iha Suai, UNAMET preokupa loos ho aktividades hosi grupu milisia Laksaur no Mahidi sira.
6. Aktividades hirak ne'e halo numeru ema refujiadu sira barak ba beibeik maka hela iha igreja Suai nian.
7. "Ami nia diskusaun liuliu ko'alia kona ba oinsa atu rezolve aktividade milisia hirak ne'e nian," Sr Martin esplika.
8. Iha Maliana, Sr Martin dehan katak mosu tiha ona hahalok seriu boot sira ne'e molok ninia grupu to'o iha ne'eba.
9. "Problema liuliu ne'e iha Bobonaro maka, hodi uluk kedas, ofisial senior sira lakohi rekonhese sira nia obrigasaun hodi autoriza grupu pro-independensia atu hala'o sira nia servisu.
10. Problema ne'e autoridades Indonezia sira nian hatene kleur ona.

Figure 2: Sentences to be translated in the experiment.

Decoding Results

Thirteen people participated in our Tetun/English human translation experiment. They included linguists, computational linguists, computer scientists, and others. Participants were free to organize themselves into teams, and two did so, leaving a total of twelve teams. The instructions, delivered over the web, were to view the Tetun test document as a code for English—and then to decode as best as possible. Participants were given 1102 simple sentence translations in a downloadable file. They were free to work by hand, to implement

computer-based tools for assistance, or to completely automate the decoding process. Computer-based tools could be held in private or distributed freely to other teams. The task time was one week; mild incentives were promised to the team with the best decodings.

One very useful keyword-in-context tool was made available. This tool accepts a word or phrase in Tetun or English and displays a list of all sentence pairs containing the word or phrase, highlighting it where it appears.

After all submissions were in, results were evaluated by two non-participant judges. Both judges were native English speakers, and they were asked to score individual sentence decodings on the basis of both accuracy and fluency, on a scale of 1 (bad) to 5 (good). Judgments were made with respect to the reference English translations that were withheld from the participants; the judges were allowed to see the original Tetun sentences, but had no Tetun competence. This kind of monolingual evaluation has its drawbacks. For example, if the reference translator simply gets it wrong, then even a great translation will be scored poorly. Or the reference translator may drop or add minor facts in her quest to produce readable text; two perfectly good translations may then disagree. However, we do not believe such problems hampered our ability to answer our basic question.

> *Some of the participants were able to produce translations that were strikingly similar to the reference translations, which we therefore assume to be reasonable translations.*

We reproduce one participant's decodings in Figure 3. Reference translations are shown in Figure 4. Evaluator judgments appear in Figure 5. See Figure 6 for sample statistical MT results. Each row stands for one participant, and each column stands for one sentence. Each cell in the matrix records the two judges' scores, separated by a colon.

Note that there was a wide range of scores. People differed significantly on how much time they devoted to the task. This is partly a function of patience—when faced with any sort of puzzle or brain-teaser, people sort themselves out according to temperament as well as skill. Also, not all the participants were native English speakers, and this affected fluency. (Note that of the top-scoring three, one was a native English speaker, and two were native German speakers). Finally, one participant entered the (lightly-edited) results of a corpus-based MT system. After the evaluation, judges were asked to identify the best and worst individual sentence translations—the former was produced by a linguist, the latter by the MT system.

We believe that if all decoders had worked as a single team, the resulting translation would have scored better than any individual translation; in our debriefings (described in the next section) we found that even the highest-scoring decoders could have easily been convinced to change their minds on various points.

1. Senior UNAMET Officials ("Functionaries") for Maliana Suai and Viqueque are scheduled for (=planning and expected to go to) Kuarta, along with Ambassador Agus Tarmidzi, President of the Indonesian "Service Force" (Tarriff Force), Brigadier General Satoris, Official of the Police (Polri) in East Timor, and several Indonesian Senior Officials.
2. Mr. Ian Martin, the [UN] Secretary General's Special Envoy (=representative) for East Timor, explained: "We go to the[se] three locations, because these locations are locations that we are very worried about."
3. Mr. Martin said that "Even though the registration process shows signs of improvement, very serious events have taken place since the one in Viqueque, because a number of people have had to give up living in their home.
4. This is a situation that we must watch/look at."
5. In Suai, UNAMET is worried about the activities by the Laksaur and Mahidi militias.
6. These activities, which have left an increasing number of people displaced, occur [mainly?] in the [area around?] Suai.
7. "Our discussion particularly deals with the question how to resolve this militia activity," Mr Martin explained.
8. In Maliana, Mr. Martin said that very serious misconduct had occurred before his group arrived (there).
9. The problems particularly in Bobonaro are, to begin with, that senior officials have refused to recognize their obligation to authorize (permit/allow) pro-independence groups to perform their function/role.
10. The Indonesian authorities have long known about this problem."

Figure 3: Decodings from one participant in the Tetun/English experiment.

1. Senior UNAMET staff went to Maliana, Suai and Viqueque on Wednesday with Ambassador Agus Tarmidzi, Chairman of the Indonesian Task Force, Brigadier General Satoris, Senior Polri Officer in East Timor and other senior Indonesian officials.
2. "We went to those three places because they are places that are of the most concern to us," Mr. Ian Martin, the Special Representative of the Secretary-General for East Timor, explained.
3. "Despite a successful registration, there has recently been serious disorder in Viqueque, and as a result a large number of people have fled from their homes," Mr. Martin said.
4. "That is a situation that urgently needs addressing," he said.
5. In Suai, UNAMET is concerned about the continued activities of the Laksaur and Mahidi militia groups.
6. These activities have led to increases in the numbers of internally displaced people in the church compound in Suai.
7. "Our discussion focussed on the need to take action to reign in the activity of those militia groups," Mr. Martin explained.
8. In Maliana, Mr. Martin said the party arrived soon after several serious incidents.
9. "The fundamental problem in Bobonaro has been, from the beginning, the refusal of senior officials to recognize their obligation to allow pro-independence groups to operate.
10. That problem in Bobonaro has been well known to the Indonesian authorities for some time."

Figure 4: Reference translations.

id	scores	avg
006	3:4 5:5 4:4 5:5 5:5 4:3 5:4 5:5 5:5 5:5	4.55
001	4:5 5:5 3:4 5:5 5:4 5:5 5:5 5:5 4:5 2:3	4.45
003	2:4 5:4 5:4 5:5 5:5 5:5 5:4 5:5 5:2 5:4	4.45
011	4:4 4:4 2:2 4:5 4:3 5:5 4:5 4:4 5:4 5:5	4.10
007	4:4 4:5 4:3 4:2 5:5 5:5 5:5 3:5 4:2 2:2	3.90
010	4:4 3:4 4:3 5:5 5:5 2:4 5:5 3:2 2:2 3:2	3.60
002	4:5 3:4 4:5 3:2 5:4 2:2 4:4 3:2 5:5 2:2	3.50
012	3:3 4:4 2:3 1:4 5:5 3:2 4:4 5:4 5:4 1:2	3.40
008	3:3 2:3 1:2 3:5 4:4 4:4 5:4 2:3 5:3 1:2	3.15
004	2:3 4:3 3:2 5:5 3:5 2:3 3:3 1:1 1:1 2:1	2.65
005	3:3 2:3 1:1 2:3 3:3 2:2 3:4 1:2 2:2 1:2	2.25
009	3:3 2:3 3:3 5:5 3:4 1:1 1:1 1:1 1:1 1:1	2.20

Figure 5: Evaluator judgments for all participants. Rows indicate participants, columns indicate sentences, and entries represent the two judges' scores.

6. Aktividades said the new documentation people displaced number journalists serious living igreja Suai.

Figure 6: Sample statistical MT results (IBM model 3).

Decoding Strategies

Here we cover strategies used by human decoders.

Left to Right, Conceptualize

The most common method was to (1) gather potential word translations for each Tetun word, moving left to right over a sentence and using the bilingual corpus, (2) pick word translations that seem to make sense with each other, (3) guess the basic idea of the whole sentence from those translations, and (4) generate a good English sentence expressing that idea. For example, from sentence 4 some decoders could come up with <*situation*, *need*, *watch*>, which could then be turned into *it is necessary to watch the situation* or *we must watch the situation*, among many other formulations we observed.

In generating English, we noticed that decoders frequently shifted part of speech. While the bilingual corpus contains many translations for the Tetun word *presiza* (*needed*, *necessary*, *need to*, *required*, *have to*), decoders felt free to go outside this list (e.g., *must*). This was particularly important for words that were observed only once or twice in the bilingual corpus. Synonym substitution was also frequent. Decoders tapped into extensive English knowledge by supplying articles, copulas, and plurals which they deemed to be missing in Tetun. In expressing the overall idea in English, final word ordering seemed to be a matter of English grammar only, but many decoders reported their discovery that Tetun adjectives often followed nouns and had to be re-ordered.

Intersections and Locality

When consulting the bilingual corpus for instances of a particular Tetun word, decoders would often scroll past long sentence pairs until they found a short one. Short sentence pairs offer a smaller list of potential translations. Given two or three short sentences with the word *ema*, for example, it was easy to look at the intersection of words in the different English translations. This intersection often tolerated inexact matches, e.g., *person/people*. This may be why most decoders preferred to do this operation by hand rather than implement matching algorithms. One decoder implemented IBM Model 1 (Brown *et al.* 1993) but found the resulting probabilistic dictionary to be fairly unhelpful.

Many words do not appear in a large number of sentence pairs. In this case, another strategy was to use locality. Consider trying to determine a translation for *liuliu* in a very long sentence pair. First, decoders easily noticed that the last few phrases often seem to be translations of one another, effectively creating a smaller sentence pair:

... grupu milisia sira liuliu iha loromonu nian iha Timor Lorosa'e.
... militia groups particularly in the west of East Timor.

Next, the phrase *grupu milisia* seems on its face to

translate as *militia groups*, while *Timor Lorosa'e* is clearly *East Timor*. Decoders were then able to hypothesize that *loromonu* translates as *west*. They could then look up this word (which was not in the task-text) and find that it occurred twice, along with the English word *western*. By process of elimination, *liuliu* could then be tentatively linked to the word *particularly*. Another scan of the bilingual corpus showed *liuliu* co-occuring with the phrase *in particular*, effectively confirming this link.

Cognates and Proper Names

To be able to apply locality, decoders needed to know at least some word translations up front. All decoders made heavy use of cognates (Tetun/English word pairs with similar spellings), as shown in the example above. Proper names can be seen as a special case of cognates, undergoing no change in spelling. Punctuation and numerals also provided important anchor points.

Unknown Words

Many words in the translation task did not appear in the bilingual corpus at all, but they still had to be translated. Decoders developed several strategies for handling them. Proper names were easiest, and could be translated without change. However, in the process of "getting the idea" of the sentence, it was important to figure out the type of entity referred to by the proper name. If the rest of the sentence was very clear, the type could be inferred, although this was not foolproof. Some decoders decided to go to web search engines as well. It was not difficult find out from the web that *Viqueque* is a town in East Timor.

Some words were unknown because of the severe lack of spelling standardization in written Tetun. It was difficult for decoders to search for spelling variants, but some were able to compile spelling variation lists while observing sentence pairs retrieved for other purposes. For example, the Tetun translation for *East Timor* appears variously as *Timor Loro sa'e*, *Timor Loro sae*, *Timor Loro-sa'e*, *Timor Lorosa'e*, *Timor Loroasa'e*, and *Timor Lorosae*.

Cognates played a very important role in translating unknown words. It proved easy to decode the phrase *aktividade milisia* even though the first word was never observed in the bilingual corpus. As mentioned above, cognates were also important for anchoring and locality. The bilingual corpus reveals a large number of Tetun/English cognates, such as *grupu/group* and *diskasaun/discussion*. However, there are a much larger number of Tetun/Portuguese cognates. Some decoders could call on friends with Portuguese knowledge to confirm hypotheses. A larger number of decoders knew Spanish, which turned out to be sufficient. For example, the word *igreja* appears in sentence 6, but never in the bilingual corpus. Only half of the decoders were able to translate this word. Several decoders noticed that *igreja* is similar to the Spanish word *iglesia*, which means *church*. One decoder used an online Portuguese/English dictionary to confirm this hypothesis.[2] Other words such as *xefe* (boss), *esplika* (explain), and *preokupasaun* (worry) could be similarly decoded.

Short Tetun Words

Many short Tetun words were easily handled by the decoders. For example, *ho* and *no* both occur frequently translated as *and* in the bilingual corpus. *Iha* seems to translate mostly as *in* or *on*, and decoders easily made this choice when generating fluent English translations.

Other Tetun words were much more difficult because they seemed to have no clear translation, even in the bilingual corpus. These included words like *maka*, *ida*, *sira*, *nian*, *tiha*, and *ona*. The decoders found these frustrating because they were quite frequent in all texts. For example, no decoder could report any theory about the meaning or function of the word *maka*—despite the fact that it occurs forty-two times in the bilingual corpus. The solution adopted by most decoders was to simply ignore these words. This is the reverse of the article/copula/plural situation described above, in which short English words seemed to have no Tetun equivalents.

Some of these puzzles were resolved when we later (subsequent to this experiment) obtained a small handbook for the Tetun language, e.g.:

> A useful word in Tetum is **maka** (often shortened to **mak**) which means 'is what', 'is the one that', 'is the thing that', e.g. **Serveja mak ami hakarak.** = Beer is what we want. (Hull 1999)

Indeed, in Tetun/English translation, it is usually more natural to ignore *maka* than to translate it. We note that *mak* occurred 124 times in our bilingual corpus, but that it did not occur in the test corpus—so decoders had little incentive to figure out that *mak* and *maka* were equivalent.

Phrases

Decoders frequently tried to look up two- and three-word phrases in the bilingual corpus. For example, *prosesu tau naran* occurred several times, translated as *registration process* or simply *registration*. As another example, *tiha ona* was observed many times, but was eventually ignored by decoders, like *maka*. It was convenient that *tiha ona* could be treated as a unit for such reasoning.

Reverse Lookup

Sentence 2 contains the word *ami*, which always co-occurs with the English word *we*, and is therefore easy to translate. Sentence 7, however, begins with the

[2]The use of this dictionary, the use of Portuguese-speaking informants, and the use of the web for semantically typing proper names made up the rare use of outside material. The highest-scoring participant used only the bilingual corpus.

phrase *ami nia*. The word *nia* seems to be quite ambiguous, co-occuring with *you*, *your*, *he*, *its*, etc. The phrase *ami nia* never occurs in the bilingual corpus. Almost all decoders were able to work out this puzzle.

The phrase *imi nia* was observed to translate as *your*, while *imi* alone translates as *you*. This was enough for decoders to hypothesize that *nia* is a possessive marker. To confirm this hypothesis, they could look up the English words *they* and *their* and find the Tetun translations *sira* and *sira nia* (note that this type of reasoning reverses the normal Tetun-based lookup process). Some decoders noticed that the English word *he* translates as *nia*; however, there is no such phrase *nia nia*, as the possessive *his* is rendered in short form *ninia*.

In this case, the decoders clearly knew what they were looking for, and could apply reasoning that was more sophisticated than co-occurrence counting.

Negation

Most decoders were able to determine that there is no separate Tetun word indicating negation. Rather, the letters *la-* are frequently prefixed to negate some item. This could be determined again by reverse lookup, i.e., determining which Tetun words co-occur with the English words *not* and *no*. As with reasoning about cognates and spelling variations, it was necessary to look at patterns within words as well as co-occurrences at the word level.

Days of the Week

Sentence 1 contains a puzzle in the phrase *Kuarta-feira*. The bilingual corpus contains the word *Sexta-feira*, translated as *Friday*, but no other instances of either *Kuarta* or *feira*. Most decoders settled on *Wednesday*, while a few picked *Thursday* or nothing at all.

Multi-Sentence Flow

Several decoders reported that it was useful to move back and forth across sentences in the text they were translating. For example, the first sentence mentions *Maliana*, *Suai*, and *Viqueque*, although it is difficult to tell what types of entities these are. However, each is covered in turn by subsequent sentences, where it becomes clear that they are best interpreted as places where things happen. Decoders also reported that the mention of militia problems (in the middle of the text) helped them interpret later passages.

Deliberate Ambiguity

In the phrase *grupu milisia Laksaur no Mahidi sira*, decoders did not agree on the semantic type of *Laksaur*. Some imagined it to be the town where the militia group was, while others imagined it to be the name of the militia group. At least one decoder was unsure, and rather than risk the translation *militia groups in Laksaur and Mahidi*, he translated the phrase ambiguously as *the Laksaur and Mahidi militia groups*.

Non-Strategies

We found that Tetun syntax did not play an important role in any of the decoders' work. No one attempted to draw Tetun parse trees as an intermediate step in decoding. We also found that no decoders drew any conclusions based on the fact that Tetun is in the Austronesian language family, and might therefore behave in certain predictable patterns.

Discussion

Our basic result shows that people can learn to translate a language they do not know if they are given a small bilingual corpus. They do so by employing a number of strategies. Corpus-based MT approaches perform badly on the same task, so there is much room for improvement.

It is interesting to consider which of the decoding strategies are amenable to being formalized in computer algorithms. We consider this to be a good topic for future exploration. Clearly, people can figure out how to translate Tetun without knowing anything about the language. But in this experiment, they knew quite a bit about English, the target language.[3] Moreover, they had some ability to synthesize a number of word translations into a coherent idea of the sentence. These are difficult areas. It would be useful for the machine to locate an appropriate concept for a Tetun word based on its several different translations. For example, the translations for *presiza* (*needed*, *necessary*, *need to*, *required*, *have to*) are enough to identify the general logical concept of necessity. Existing natural language generation programs such as Nitrogen (Langkilde & Knight 1998) can then render this general logical concept in many ways, depending on the other phrases generated in the same sentence. For example, the input:

```
(n / NECESSITY
  :domain (o / OBSERVE
             :patient (s / SITUATION)))
```

is rendered by Nitrogen automatically in over 66,000 ways; relatively highly-ranked ones include:

> The situation must be watched attentively.
> A situation must be watched attentively.
> It must be watched attentively that it is situated.
> It is necessary to watch attentively this situation.

At the very least, a machine should have some basic capability to expand the set of translations beyond those observed in the corpus. In our machine experiments, we frequently observed our program struggling

[3] It would be extremely interesting to run such experiment between two unknown languages. Target language considerations such as word order would have to be made on the basis of observed target-language corpora only. (Knight 1997) contains a small artificial corpus along these lines, between imaginary languages Centauri and Arcturan. These languages turn out to be English and Spanish in disguise, allowing human decoders to mentally simulate statistical algorithms without bias.

with a word like *saw* when it needed a word like *sight*. Expanded translation sets can be built from inflectional morphology, derivational morphology, synonym-finding, and other processes.

We carried out some MT experiments using two of the human strategies described above. These were strategies for dealing with cognates and short function words.

We found that standard statistical word-alignments between Tetun and English training sentences were quite inaccurate. Obvious cognate pairs were not connected, as the training algorithm made no use of word-internal features. One easy way to address this problem is to supply a list of cognate pairs as an additional "corpus" appended to the real corpus. This biases the word-alignments search in favor of connecting cognate pairs. We first generate a cognate-pair candidate for each word co-occurrence in the training corpus. For example, from a sentence pair of length n, we will list out n^2 candidates. Most of these candidates are not translations at all, so we restrict candidate pairs to begin with the same letter. We then use an algorithm described by Noah Smith in (Al-Onaizan *et al.* 1999) to find spelling similarities and simultaneously rank the candidate list. We take the top of this list as our cognate corpus, which we append to the regular training corpus. Here are some of the cognates suggested automatically:

0.514 problema/problem	0.494 promove/promote
0.496 prova/prove	0.492 proposta/proposal
0.496 imparsial/impartial	0.489 forma/forms

Because these word-pairs not only look alike but also co-occur in the corpus, they are fairly reliable. Further down the ranked list, false cognates begin to appear, e.g.,

0.302 fila/fear 0.302 pessoal/personnel

In inspecting our word alignments, we also found that most of the Tetun function words were connected to various English words. Given a larger corpus, we expect that the training algorithm would learn to generate Tetun function words from the special NULL token, effectively telling the automatic decoder not to hypothesize English translations for them. However, our decoder does not make such hypotheses, loading up English translations with lots of extra words. We do not yet know how to automatically identify such NULL-generated words with a small parallel corpus, but we can make use of our manual analysis. Prior to decoding, we simply remove all "stop-words" (e.g., *maka*) from any Tetun document. We have found that the translations improve when we do this.

For example, before working with cognates and function words, our automatic translation of sentence 4 was:

> This is serious situation which was necessary put.

Afterwards, our translation was:

> Situation which need to see.

Of course, there is a great deal of the difference between our machine translations and those of human decoders, and we believe it is worth continuing along these lines.

Finally, we note again that our human decoding experiments were run over the web—we provided training and testing corpora, search tools, and evaluation software. This facility is open to the public,[4] and we hope to add other languages. We believe it should be of educational value in computational linguistics, artificial intelligence, and linguistics curricula.

Acknowledgments

We would like to thank Katya Shuldiner for collecting and collating human decodings. Many thanks to David Purdy and Richard Whitney for doing evaluation, and to participants Jonathan Gratch, David Lugo, Franz-Josef Och, Andrew Philpot, Bonnie Glover Stalls, and Marcelo Tallis. This work was supported in part by DARPA-ITO award N66001-00-1-9814.

References

Al-Onaizan, Y.; Curin, J.; Jahr, M.; Knight, K.; Lafferty, J.; Melamed, D.; Och, F.; Purdy, D.; Smith, N. A.; and Yarowsky, D. 1999. Statistical machine translation, final report, JHU Workshop 1999. Technical report, CLSP, Johns Hopkins University.

Alshawi, H.; Buchsbaum, A.; and Xia, F. 1997. A comparison of head transducers and transfer for a limited domain translation applications. In *Proc. ACL.*

Brown, P.; Pietra, S. D.; Pietra, V. D.; and Mercer, R. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics* 19(2).

Church, K. 1993. A program for alinging parallel texts at the character level. In *Proc. ACL.*

Hull, G. 1999. *Tetum: Language Manual for East Timor.* Academy of East Timor Studies.

Knight, K. 1997. Automating knowledge acquisition for machine translation. *AI Magazine* 18(4).

Langkilde, I., and Knight, K. 1998. Generation that exploits corpus-based statistical knowledge. In *Proc. COLING/ACL.*

Melamed, D. 2000. *Empirical Methods for Exploiting Parallel Texts.* MIT Press.

Nagao, M. 1984. A framework of a mechanical translation between Japanese and English by analogy principle. In Elithorn, A., and Bernerji, R., eds., *Artificial and Human Intelligence.* North-Holland.

Och, F.; Tillmann, C.; and Ney, H. 1999. Improved alignment models for statistical machine translation. In *Proc. EMNLP/WVLC.*

Wu, D. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics* 23(3).

[4] `www.isi.edu/natural-language/mt/contest/`