Toward a Science of Machine Translation

Francis Bond
NTT Machine Translation Research Group
NTT Communication Science Laboratories
Nippon Telephone and Telegraph Corporation
2-4 Hikari-dai, Seika-cho, Soraku-gun, Kyoto, 619-0237, JAPAN
bond@cslab.kecl.ntt.co.jp

1   Introduction

The fact that machine translation output does not reach the level of good human translations is well known. However, there has been surprisingly little attention paid by the machine translation community to how humans achieve such results.

In this paper, I suggest several ways to improve machine translation, based on the best practices of human translators, as described in Nida's (1964) *Toward a Science of Translating*. I call this approach multi-pass machine translation (MPMT), as it crucially relies on processing the text more than once. It is similar to the opportunistic *bricoleur* approach of Gdaniec (1999) in that it sets out to use the means at hand, adding to or changing them as necessary. As Schütz (2001) points out, much of the research in the past decade has concentrated on the important but non-core issues of integrating MT into DTP formats and HTML. In this paper I concentrate on improving the MT engine itself. The resulting approach integrates much recent research into a single system.

2   Toward a Science of Machine Translation

Nida (1964:246-247) sets out the following nine steps to be employed by a competent translator (with some steps omissible):

1. Reading over the entire document

2. Obtaining background information

3. Comparing existing translations of the text [if they exist]

4. Making a first draft of sufficiently comprehensible units

5. Revising the first draft after a short lapse of time

6. Reading aloud for style and rhythm

7. Studying the reactions of receptors by the reading of the text by another person [omissible]

8. Submitting a translation to the scrutiny of other competent translators [omissible]

9. Revising the text for publication

In the following nine subsections, I will suggest how these can be adopted by an MT system. I will concentrate on the translation of texts; interpretation is a different problem which requires different solutions. I will then combine them to produce a multi-pass machine translation system (§ 3).

2.1   Reading over the entire document

Before translation, a system should parse the entire document, in order to do the following things:

- Identify the source language (possibly per segment)
- Identify the domain
- Identify terms and named entities

Knowledge of the source language is an obvious prerequisite for machine translation, as it enables the system to select the appropriate grammar and lexicons. In many settings, this knowledge can be assumed, but the language is not normally identified on the web. Another problem that requires language identification is the reasonably common practice of interspersing text with text of one or more different languages.

Knowledge of domains has been shown to improve the quality of translation in at least two studies. Yoshimoto et al. (1997) using 41 hierarchically organized domains (which they call Subject Areas) marked on 9,000 words, were able to improve the translations of 12% of the badly translated nouns ($\frac{46}{383}$). Lange & Yang (1999) used 77 Domains (TERMinology CATegories) and 30 Topical Glossaries. In a first pass through they chose two appropriate domains (the domains with the greatest number of domain-tagged words). With these domains set, there were changes in 0–40% of translations, and the majority of the changes were improvements.

Identification of terms and named entries are both robust research fields in their own right. It is important for a translator to identify them, so that they can ensure they have the correct translations for that domain, as discussed in the next section.

2.2   Obtaining background information

Once the domain is known, and lists of terms and named entities have been produced, the system must try and find appropriate translations. This is often the most time-consuming task for human translators, especially for new domains and unknown proper nouns.

Identifying unknown proper nouns has been shown to give a 20.8% improvement in quality of translation of proper nouns (Yoshimoto et al. 1997). Knowing whether something is a proper noun is crucial to both the analysis of the text and the selection of the translation. Further, Kim et al. (2001) showed that creating a user dictionary of unknown words in a first pass, increased the accuracy of the morphological analysis for Korean in a second pass by 2.26%.

Translations of unknown words can be extracted from unaligned corpora (Tanaka & Matsuo 1999), or from the world wide web (Grefenstette 1999).

Systems that use stochastic models of source and target text, such as n-grams, or rankings for choosing translations and variations, should retrain on, or adapt to text of similar domains for their language models.

2.3 Comparing existing translations

One proof of the utility of comparing with existing translations is the well-attested benefits of the use of translation memories (Planas 1999). Their value is so widely known that there is little need to say more. If existing translations are available they can also be used to extract rules (Yamada et al. 1995), or provide data for example-based machine translation. Further, a translation can help to guide a system's understanding of the source text by adding extra constraints on its parse, in the same way that a human may look to a translation to clarify the meaning of a section they find hard to understand. Wu (1997) shows a way of exploiting a translation to parse a text using an inversion transduction grammar.

2.4 Making a first draft

The actual translation process itself should use a good rule-based system with the correct domain dictionaries and unknown words and their translations identified and put into a user dictionary, possibly combined with a translation memory to translate previously encountered text. This creates the first draft.

2.5 Revising the first draft

There have been some suggestions to check for deceptive cognates (Isabelle et al. 1993), or omissions in translation (Russell 1999), and to choose articles (Chander 1998). At present, these techniques are not yet so useful in practice, probably because the revision systems are not able to consider the meaning of the text. Naruedomkul & Cercone (1997) suggested an MT architecture that would first translate and then revise poor translations but they did not implement the revision part of the system.

One task that could be considered a revision is the addition of translator's footnotes or notes. Whether they are appropriate or not depends on the target audience. A translator's note is useful if a word without a good translation equivalent, but with a known definition, appears several times. In this case it should be glossed with a definition (the note) the first time and then transliterated in subsequent uses (possible in a different font). For example:

> There were several *akah* (a kind of turtle) lying on the beach. . . . The next morning, I was awakened by an *akah* falling on to my sleeping bag.

It is also common for named entities to be introduced by an explanation in some reporting. For example, in non-Japanese English-language newspapers, NTT is often introduced with an explanation such as *NTT, Japan's dominant telephone company,* . . . . This is not necessary in a Japanese text, as everyone presumably knows that NTT is Japan's dominant telephone company, but becomes necessary in translation.

It would be useful to add such notes either during the translation, as part of an initial revision, or as part of the final revision before publishing.

2.6 Reading aloud for style and rhythm

The style and rhythm of a text are analogous to its probability: if it sounds natural then it should be a likely sentence, and if it is an unlikely sentence it should sound unnatural. I take this step to suggest a translation model which generates all possible **variations** (translations with almost identical semantic restrictions) and chooses the best one based on a target language model. This is similar to the system proposed by Knight et al. (1994), who use n-grams as the target model.

2.7 Studying the reactions of other receptors

In order to decide which parts of the translation are of high quality, the text should be parsed by a target language parser, and scored in some way. Bernth (1999) has shown that a confidence index is useful when presenting text to end users, and suggested a score based on the translation process. This is a good idea, but requires detailed knowledge of the transfer process. A simple monolingual metric measured on the target text is more robust.

2.8 Submitting a translation to other translators' scrutiny

By using a confidence measure on the output text, it is possible to rank the outputs of more than one translation system. Callison-Burch & Flournoy (2001) trained a trigram language model on 2,000,000 words of English and used it to choose among the outputs of four Japanese-English systems. This gave an overall quality of 74% compared to 70%, 58%, 40% and 27% for the individual systems. Choosing between two French-English systems gave even better results: 84% compared to 76% and 56%. From this we can see that it is well worth consulting other translators, should you have access to them.

2.9 Revising the text for publication

Finally, to prepare the text for publication, it is important to restore any markup. To do this the system must keep a record of source text equivalences and tags, as well as any text-style tags.

3 Multi-Pass Machine Translation

In this section I will describe two versions of a the Multi-Pass Machine Translation system. The first is a single engine system, the second is a multi-engine system, that is, a system that combines the outputs of more than one system.

3.1 Single Engine Multi-Pass Machine Translation System

The basis of a single engine Multi-Pass Machine Translation system is a semantic transfer system. The text is analysed with a rule-based parser, and the best parse chosen using a stochastically trained source language model. The output should be a semantic representation that is as language independent as possible. In practice, a

full semantic analysis is an AI complete problem, and thus far from being solved. For the time being, we must resort to Multi-Level Transfer (Ikehara et al. 1991; Ikehara et al. 1996) where a text is analysed as far as possible, and then transfered to the target language. The transfer stage will also have a rule-based core with a transfer model to chose between alternatives. Finally, the target language text is generated from the target language semantic representation. In order to benefit from work in mono-lingual parsers, generators and lexicons, it is desirable to keep the analysis and generation systems as modular as possible. The enhancement of this system by the multi-pass model is shown in Figure 1. Although here I assume that the translation is done sentence by sentence, as all the current systems I know of do, the architecture is equally applicable to a discourse-based system such as that proposed by Marcu et al. (2000).

---

1. Identify the source text language (if unknown)

    - Check the language of each unit (sentence or paragraph)

2. Pre-parse the text

    (a) Extract named entities
        Put named entities into local dictionary
    (b) Run a morphological analyser to identify any unknown words
        Add unknown word candidates to local dictionary
    (c) Identify the genre and domain(s)

3. Find translations for the named entities and unknown words
   Enter them into local transfer and target dictionaries

4. Train the source and target language target models using text from similar genre and domain(s)

    - If there are any existing translations, train the transfer rules on them

5. Translate using the local dictionaries, appropriate domain dictionaries, and the trained source, transfer and target models

6. Check the style of the resulting text using a target language checker

7. Restore any markup from the source text.

    - Possibly add explanatory notes or footnotes

Figure 1: Single Engine Multi-Pass Machine Translation

The multi-pass method relies critically on two things. The first is having enough processing power and training data to be able parse several times, and retrain models on the fly. Although this is only just becoming available, it is part of the general trend

to push more grunt work onto the computer, so that humans can do other things. The other thing is an integrated mixture of rules and stochastic rankings, which I see becoming more and more dominant in natural language processing in general. In this sense I agree with Och & Ney (2001), although I would like to incorporate statistical models into a rule-based system, rather than add linguistic knowledge to a statistical model.

I hope that this architecture, and the advances in natural language processing it is based on, will help to bring machine translation ever closer to the capabilities of human translators.

3.2 Multi-Engine Multi-Pass Machine Translation System

At the current level of success of machine translation systems, a multi-engine Multi-Pass Machine Translation system along the lines suggested by Callison-Burch & Flournoy (2001) seems worth building. This system translates a text with multiple translation systems, and then chooses the best output for each sentence. The passes could in fact be carried out in parallel, even on different machines. Callison-Burch & Flournoy (2001) used n-grams to select the best result from several rule-based systems. This could be a problem if a statistical system was included, as its results will always be smooth, thus an n-gram-based ranking would be biased towards its results. To choose fairly, the ranking would have to be based on fidelity as well as fluency.

However, for really high quality translation (equalling the capabilities of a good human translator), translating one sentence with one system and one sentence with another is not a good strategy. A good text must be coherent, written in the same register, with reference to entities depending on how they have been referred to in the past. It is not possible to do this by combining the output of multiple systems sentence by sentence. For the same reason, translation memories can not be expected to provide the best translations of complete texts, unless the whole text is in the memory. That is not to deny that they are extremely useful now, due to the poor performance of current machine translation systems.

Therefore, rather than combining the results of multiple systems as described above, in a perfect world a single-engine MPMT system would instead combine all of the knowledge (rules and lexicons) used by the multiple systems. This would be enhanced by any transfer rules and statistics it could learn from existing translations, and the result used to produce a single coherent translation. This is the closest to Nida's (1964) strategy. The main problems to doing this are more social than technical. This is not to belittle the genuine technical problems involved with combining multiple knowledge sources, but rather to acknowledge that they are dwarfed by the legal problems of gaining access to the raw linguistic resources used in multiple machine translation systems.

4 Conclusion

As processing power has increased it is becomingly more and more feasible to process a text more than once when translating. This is how the best human translators produce their translations, and provides a way for machine translation systems to improve their

quality. I therefore propose a new approach to lead us on to the goal of high quality fully automatic machine translation: Multi-Pass Machine Translation.

Postscript

The inspiration for this paper came from my own experiences as a free-lance translator. When I had to translate something, I would always first try to read some other source and target language texts from the same domain, so that I had a rough idea of what things meant, and what the technical terms were. I would then write down all the words I didn't understand, and try and find their translations. If I couldn't find a translation I would even go as far as to ask the authors if they knew the translations. Without doing this, I couldn't make a decent translation. When I watched professional translators in a Japanese newspaper company doing their jobs and then read Nida (1964), I realised that I wasn't alone in doing this, and wondered whether a similar process could improve a machine translation system as well. The result is this paper.

Acknowledgments

References

Bernth, Arendse: 1999, 'A confidence index for machine translation', in *Eighth International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-99*, Chester, UK, pp. 120–125.

Callison-Burch, Chris & Raymond S. Flournoy: 2001, 'A program for automatically selecting the best output from multiple machine translation engines', in *Machine Translation Summit VIII*, Santiago de Compostela, pp. 63–66.

Chander, Ishwar: 1998, 'Automated postediting of documents', Ph.D. thesis, University of Southern California, Marina del Rey, CA.

Gdaniec, Claudia: 1999, 'Using MT for the purpose of information assimilation from the web', in *Workshop on Problems and Potential of English-to-German MT systems*, TMI, Chester, UK, `http://www.research.ibm.com/people/g/cgdaniec/tmi99.html`.

Grefenstette, Gregory: 1999, 'The WWW as a resource for example-based MT tasks', in *Translating and the Computer 21: ASLIB'99*, London.

Ikehara, Satoru, Satoshi Shirai & Francis Bond: 1996, 'Approaches to disambiguation in **ALT-J/E**', in *International Seminar on Multimodal Interactive Disambiguation: MIDDIM-96*, Grenoble, pp. 107–117.

Ikehara, Satoru, Satoshi Shirai, Akio Yokoo & Hiromi Nakaiwa: 1991, 'Toward an MT system without pre-editing – effects of new methods in **ALT-J/E**–', in *Third Machine Translation Summit: MT Summit III*, Washington DC, pp. 101–106, (`http://xxx.lanl.gov/abs/cmp-lg/9510008`).

Isabelle, Pierre, Marc Dymetman, George Foster, Jean-Mard Jutras, Elliot Macklovitch, François Perrault, Xiabao Ren & Michel Simard: 1993, 'Translation analysis and translation automation', in *Fifth International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-93*, Kyoto, pp. 201–217.

Kim, Seonho, Mansuk Song & Yuntae Yoon: 2001, 'Proper analysis of unknown words using local dictionary', in *19th International Conference on Computer Processing of Oriental Languages: ICCPOL-2001*, Seoul, pp. 439–444.

Knight, Kevin, Ishwar Chander, Matthew Haines, Vasileios Hatzivassiloglou, Eduard Hovy, Masayo Iida, Steve K. Luk, Akitoshi Okumura, Richard Whitney & Kenji Yamada: 1994, 'Integrating knowledge bases and statistics in MT', in *Proceedings of the 1st AMTA Conference*, Columbia, MD.

Lange, Elke D. & Jin Yang: 1999, 'Automatic domain recognition for machine translation', in *Machine Translation Summit VII*, Singapore, pp. 641–645.

Marcu, Daniel, Lynn Carlson & Maki Watanabe: 2000, 'The automatic translation of discourse structures', in *The 1st Meeting of the North American Chapter of the ACL: ANLP-NAACL-2000*, Seattle, pp. 9–17.

Naruedomkul, Kanlaya & Nick Cercone: 1997, 'Steps toward accurate machine translation', in *Seventh International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-97*, Santa-Fe, pp. 63–75.

Nida, Eugene A.: 1964, *Toward a Science of Translating*, Leiden, Netherlands: E. J. Brill.

Och, Franz Joseph & Hermann Ney: 2001, 'What can machine translation learn from speech recognition', in *MT 2010 — Towards a Road Map for MT*, Santiago de Compostela, MT Summit VIII Workshop, pp. 26–31.

Planas, Emmanuel: 1999, 'Formalizing translation memories', in *Machine Translation Summit VII*, Singapore, `http://www.kecl.ntt.co.jp/icl/mtg/members/planas/MTSummitVIIconf.ps`.

Russell, Graham: 1999, 'Errors of omission in translation', in *Eighth International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-99*, Chester, pp. 128–138.

Schütz, Jörg: 2001, 'Blueprint for MT evolution: Reflection on "Elements of Style"', in *MT 2010 — Towards a Road Map for MT*, Santiago de Compostela, MT Summit VIII Workshop, pp. 9–13.

Tanaka, Takaaki & Yoshihiro Matsuo: 1999, 'Extraction of translation equivalents from non-parallel corpora.' in *Eighth International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-99*, Chester, UK, pp. 109–119.

Wu, Dekai: 1997, 'Stochastic inversion transduction grammars and bilingual parsing of parallel corpora', *Computational Linguistics*, **23**(3): 377–403.

Yamada, Setsuo, Hiromi Nakaiwa, Kentaro Ogura & Satoru Ikehara: 1995, 'A method of automatically adapting a MT system to different domains', in *Sixth International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-95*, Leuven, pp. 303–310.

Yoshimoto, Yumiko, Satoshi Kinoshita & Miwako Shimazu: 1997, 'Processing of proper nouns and use of estimated subject area for web page translation', in *Seventh International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-97*, Santa Fe, pp. 10–18.