Customizing Complex Lexical Entries for High-Quality MT

Rémi Zajac
SYSTRAN Software, Inc.
9333 Genesee Avenue
Plaza Level, Suite PL1
San Diego, CA 92121
USA
zajac@systransoft.com

Elke Lange
SYSTRAN Software, Inc.
9333 Genesee Avenue
Plaza Level, Suite PL1
San Diego, CA 92121
USA
lange@systransoft.com

Jin Yang
SYSTRAN Software, Inc.
9333 Genesee Avenue
Plaza Level, Suite PL1
San Diego, CA 92121
USA
jyang@systransoft.com

Abstract

The customization of Machine Translation systems concentrates, for the most part, on MT dictionaries. In this paper, we focus on the customization of complex lexical entries that involve various types of lexical collocations, such as sub-categorization frames. We describe methods and tools that leverage existing parsers and other MT dictionaries for customization of MT dictionaries. This customization process is applied on large-scale customization of several commercial MT systems, including English to Japanese, Chinese, and Korean.

1 Introduction

1.1 Background

Customization of MT systems is a problem that is only starting gain awareness. Typically, customization is reduced to the (manual) development of a simple domain-specific user dictionary that takes priority over the main dictionary of an MT system. Complex lexical entries, for example intricate sub-categorization patterns, are excluded; a fortiori, syntactic customization is excluded too.

Most previous work on automated customization used a parallel corpus, for example Yamada et als. (1995) and Su et als. (1995, 1999). Of course, example-based systems may be considered fully customized systems (Richardson et als. 2001, Pinkham et als. 2001).

Yamada et als. (1995) present a method to adapt a rule-based MT system to a new domain by using aligned sets of sentences. The method involves the comparison of the MT parse tree (presumably after transfer) with the parse tree of the manually produced translation. However, a side effect is the automatic generation of either bilingual dictionary entries or transfer rules. The interest of the method is not clear since the technical description is rather vague. In addition, there has not been any discussion on the influence of the bilingual corpus on the quality improvement. The method seems to be implemented only for simple bilingual lexical equivalences.

Su et als. (1995, 1999) suggest that customizing an MT system can be reduced to learning probabilistic parsing parameters, used to select the best parse of a non-deterministic parser. The best parse is the one that offers a translation that is closest to the manually translated sentence (or the one which produces a parse tree that is closest to the parse tree of the manually translated sentence, the paper is unclear on this point). The method does not seem to be implemented.

Current ongoing approaches based on large parallel corpora that provide the highest quality results to fully automatic customization use example-based techniques built on a substrate of a comprehensive rule-based system, as in the MSR-MT project (Richardson et als. 2001, Pinkham et als. 2001). In this approach, there is no distinction between lexical and syntactic customization. What is learned, in essence, is a set of lexicalized transfer rules that may cover entire sentences.

1.2 Customization for High-Quality MT

Our approach embeds specific customization tasks in a comprehensive approach to high-quality MT that covers:

1. The definition of a detailed and linguistically relevant document type definition, used to select the most appropriate translation parameters depending on the content of the specific XML elements (and not just the whole text).
2. The construction of a custom user dictionary to cover all domain terminology and all domain-specific lexical collocations.
3. The customization of the MT engine rules (parsing, transfer, and generation rules) to account for idiosyncrasies in style (and possibly missing syntactic constructions).
4. The use of Controlled Language to eliminate lexical problems (spelling errors, missing lexical items, inconsistency in the use of acronyms and abbreviations, etc.), and normalization of the writing style (so that the style parameters implemented in the MT system closely match the writing style actually used in documents).

MT customization is based on large-scale corpus analysis and exploitation, and continued rigorous translation quality testing. This approach to MT customization includes:

- Detailed corpus profiling and translation quality evaluation in order to derive a quantitative customization work plan that specifies how to close the gap between out-of-the-box translation quality and targeted translation quality.
- Customization tasks in a staggered fashion:
  - Definition of a translation stylesheet (for XML documents);
  - Corpus terminology (mostly nouns);
  - Lexical collocations (mostly predicates);
  - Tuning of parsing, transfer, and generation rules.

Lexical customization is typically done in two steps:

- Large scale term extraction, translation and coding;
- Manual tune-up using specialized translation quality review tools.

The construction of a custom user dictionary covering nominal terminology is, in size, the most important customization task. Once basic terminology is covered, it becomes possible to accurately extract lexical collocations between the corpus terms.

The rest of the paper describes the processes of extracting lexical collocations (Section 2), and of creating customized complex bilingual lexical entries (Section 3). We conclude on the benefits provided by a complex custom translation dictionary.

2 Extraction of Lexical Collocations for MT

Most lexical collocation extraction work is based on either extended regular expressions or robust partial parsing (e.g., Smadja 93, Debili 82). In this extraction process, we use the parser of the MT system itself to tag the corpus. Lexical collocations are then extracted by matching a list of syntactic relationships identified by the parser.

The English parser used in the process is a large-scale parser that includes over 5,000 rules and has a dictionary of over 300,000 entries. It was developed by a team of several linguists over 15 years ago, and is used for the English-Japanese, Chinese, and Korean systems mentioned in this paper, as well as in several other systems.

The tagged corpus includes syntactic and semantic relationships between heads of phrases. Among the relationships identified by the parser, we extract patterns that typically help to improve the translation of sub-categorization frames. Correction identification and translation of sub-categorized complements help to solve several issues at once:

- Attachment of prepositional phrases in English;

- Ordering of complements on the target side;
- Translation of prepositions by specific case markers (in Japanese for example).

The parser distinguishes between surface syntactic relationships and deep semantic relationships. These relationships are established using a set of heuristics that evaluate lexical constraints coded in the dictionary. For example, in order to determine whether prepositional phrases are to be recognized as an Object, the heuristics may examine a variety of clues that include:

- Does the verb entry include a sub-categorized prepositional phrase for the particular preposition?
- Does the prepositional phrase immediately follow the verb?
- Does the preposition preferentially attach to a verb?
- Does the preposition mark time or space?
- If another noun-preposition or prepositional phrase between the verb and the prepositional phrase exists, is the preposition preferentially attached to a verb or a noun?
- Etc.?

In order to achieve the greatest coverage for complex lexical expressions, we use the tagged semantic relationships only. In the following example, the syntactic object 'password' is tagged as an Object:

- *They changed the password.*

And it is tagged as a semantic Object too in the following cases:

- *They changed the password.*
- *The password was changed.*
- *The password changed by the administrator.*
- *The changed password.*
- *Passwords are changed when needed.*
- *Changing passwords is important.*

Similarly, the "Subject-Predicate" relationship marked is the Agent-Action and encompasses various surface syntactic manifestations.

There are about a dozen lexically relevant relationships, including for example:

Relationship	**Extracted instance**	**From sentence**
Verb-Object	configure <bridging>	How do I configure bridging on ARM ?
Verb-Object-Preposition	specify <direction> (in)	The direction must be specified in later software releases.
Verb-Object-Infinitival	configure <client> <obtain>	...the client is configured to obtain an IP address
Verb-Particle-Object	find out <number>	How do I find out the number of files that a process has open?
Verb-Preposition-Object	refer (to <code>)	For more details refer to the debug codes.
Noun-Preposition-Noun	configuration (for <authentication>)	Configurations for login authentication.
Adjective-Preposition-Noun	available (to <customer>)	available to end users and customers
Adjective-Preposition	equivalent (to)	is equivalent to:

In order to provide an idea of the number of such relationships that could be found in a corpus, the following table shows the number of relationships automatically extracted from a technical corpus of 5 millions words:

Type	**Size**
Verb-Object	10,569
Verb-Object-Infinitival	2,290
Verb-Particle-Object	90
Agent-Verb	4,027
Verb-Preposition-Noun	940
Noun-Preposition-Noun	1,839
Adjective-Preposition-Noun	145

3 Customization of Complex Bilingual Dictionary Entries

Extracted terms are processed in several steps:

Lexical patterns are reviewed on a monolingual (source) basis to weed out obvious parsing errors. This initial review is done very quickly.

1. A second bilingual pass identifies lexical items with potential translation problems.
2. A third pass corrects the lexical entries according to the problems exhibited in the translation examples.

Bilingual lexicographers who review lexical entries that correspond to an instance of a single relationship also perform the second translation review. These instances include the sentence from which the instance of the relationship was extracted together with the translation that was generated automatically. When the translation is wrong, the lexical entry is flagged for coding.

Verb-Object

- Support platform
- サポート アクセス・サーバプラットフォーム
- This early deployment release supports the server platform and replaces the deferred 12.a release.

Verb-Object-Infinitival

- See date view
- 見 発売 予定日 表示
- To view the projected release date of the software releases see the Software Product Bulletin.

Verb-Object-Preposition

- Base information on
- 基づ ステート情報 に
- This product has better scaling properties than an ATM because its state information is based on the topology of the virtual networks.

Verb-Agent

- Arrive packet
- 着 すべてのパケット
- Even on a virtual node configured with AFCD, all packets that arrive at a virtual node when the average queue size is above the queue limit are tail dropped.

The coding of complex lexical entries is performed by bilingual lexicographers specialized in the terminology domain, on both the source and the target parts. Depending on the type of relationship on the source part, an entry may contain all lexical elements already present in an instance (in cases of strong collocations) or be generalized. For example, a series of Verb-Object instances may be generalized to the common semantic category of the Object and the semantic category specified as a constraint on the Object. In other cases, the syntactic type may instead be specified. For prepositional objects, no constrain other than the occurrence of the preposition itself may be specified. On the target part, the default translation may be corrected; for example the translation of the preposition into the correct case marker in Japanese.

In the following Verb-Object examples, the English object is given in the entry for informational purposes only, but is not part of the lexical entry. The Japanese indicates only the translation of the verb and the case particle to be used for the Object.

On this example, the translation of the verb itself is wrong (the Japanese case particle for the Object is indicated in between parentheses):

```
<term>
      <en>contain (space)</en>
      <pos>verb</pos>
<ja citation="含む">(を)含む</ja>
      <ex>" " contains a
space.</ex>
</term>
```

It is to be corrected as:

```
<term>
      <en>contain (space)</en>
      <pos>verb</pos>
      <ja citation="はいる">(が)はいっ
ている</ja>
      <ex>" " contains a
space.</ex>
</term>
```

In the following example, the English active should be translated as Japanese passive:

```
<term>
      <en>contain (step)</en>
      <pos>verb</pos>
      <ja citation="含む">(を)含む
</ja>
      <ex>Contains the necessary
steps for employees to assist in
      customer upgrades.</ex>
</term>
```

After revision:

```
<term>
      <en>contain (step)</en>
      <pos>verb</pos>
      <ja citation="記載される">(が)記
載されている</ja>
      <ex>Contains the necessary
steps for employees to assist in
      customer upgrades.</ex>
</term>
```

The ratio of coded entries to the extracted instances varies according to: the type of relationship, coherent with the gap between the source and target languages; the distance between the corpus style and content; and the average document for which the out-of-the-box system was optimized. In the case of the highly technical corpus mentioned above, about 88% of extracted instances are coded for the English-Japanese language pair and for Verb-Object relationships.

4 Conclusion

We use a version of the SAE J2450 Translation Quality Metric modified for the evaluation of MT systems. For highly technical texts, varying a great extent upon a specific corpus (relative amount of specialized terminology, frequency of particular syntactic constructions, complexity of the writing style, correctness of the language), our out-of-the-box MT systems rank between 40-60% quality based on this metric. After initial terminology work is completed (covering basic nominal terminology), we can expect an increase of 10 to 35% in quality, with most systems reaching a quality level of 65-80%. Additional customization work on lexical collocation as described above may bring further improvements of 5-15%.

Any further customization work follows the law of diminishing returns. Depending on the type of text, addressing the following items may typically bring between 5 and 10% increase in translation quality:

- Use of detailed XML document structure in which XML elements are unambiguously associated with specific linguistic properties via a translation stylesheet.
- Customization of MT engine rules (parsing, transfer, and generation) to accommodate for the difference in frequency of some syntactic constructions (style), and occasionally the addition of new syntactic constructions.
- Use of Controlled Language that provides integrated spell-checking and promotes a consistent and simple writing style.

5 References

Jacquemin, Christian. 2001. Spotting and Discovering Terms trough Natural Language Processing. The MIT Press.

Lalaude, Myriam, Veronika Lux, Sylvie Regnier-Prost. 1998. “Modular controlled language design”. CLAW-98, Pittsburgh, PA. Pp103-113.

Pinkham, Jessie, Monica Corston-Oliver, Martine Smets, Martine Petterano. 2001. “Rapid assembly of a large-scale French-English MT system”. MT Summit VIII, Santiago de Compostela, Spain. Pp277-282.

Richardson, Stephen, William Dolan, Arul Mezenes, Jessie Pinkham. 2001. "Achieving commercial-quality translation with example-based methods". MT Summit VIII, Santiago de Compostela, Spain. Pp293-298.

Senellart, Jean, Boitet, Christian, Romary, Laurent, 2003. "SYSTRAN New Generation: The XML Translation Workflow". MT Summit IX, New Orleans, Louisiana, USA.

Senellart, Jean, Yang, Jin, Rebollo, Anabel, 2003. "SYSTRAN Intuitive Coding Technology". MT Summit IX, New Orleans, Louisiana, USA.

Senellart, Jean, Peter Dienes, Tamas Varadi. 2001a. "New generation SYSTRAN translation system". MT Summit VIII, Santiago de Compostela, Spain. Pp311-316.

Senellart, Jean, Mirko Plitt, Christophe Bailly, Francoise Cardoso. 2001b. "Resource alignment and implicit transfer". MT Summit VIII, Santiago de Compostela, Spain. Pp317-324.

Smadja, F. 1993. Retrieving collocations for text: Xtract. Computational Linguistics 19/1, pp143-177.

Smadja, F., K.R. McKeown, V. Hatzivassiloglou. 1991. Translating collocations for bilingual lexicons: a statistical approach. Computational Linguistics 22/1, pp1-38.

Su, Keh-Yih, Jing-Shin Chang. 1999. "A customizable, self-learnable parameterized MT system: the next generation". MT Summit VII, Singapore. Pp182-190.

Underwood, Nancy L., Bart Jongejan. 1999. "Profiling Translation Projects". TMI-99, Chester, England. Pp139-149.

Yamada, Setsuo, Hiromi Nakaiwa, Kentaro Ogura, Satoru Ikehara. "A method for automatically adapting an MT system to different domain". TMI-95, Leuven, Belgium. Pp303-310.