Challenges in Automated Elicitation of a Controlled Bilingual Corpus

Katharina Probst, Lori Levin
Language Technologies Institute, Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213, USA
{kathrin,lsl}@cs.cmu.edu

Abstract

In this paper we will address an uncommon but important approach to automated learning for MT, namely learning of translation rules from carefully elicited sentences. The approach is uncommon for good reason — anyone who has tried linguistic field work knows that elicitation will go awry if not carefully monitored by a human. We will address eight challenges of automated elicitation and discuss their solution in the AVENUE machine translation project. The elicited sentences in AVENUE are used to semi-automatically infer transfer rules for the desired language pair.

1 Introduction

In recent years, much of machine translation research has focused on two issues: portability to new languages, and developing machine translation systems rapidly. Since human expertise on rare languages may be in short supply, and human development time can be lengthy, automated learning of statistics or rules has been critical to both language portability and rapid development. Automated methods have typically been trained on uncontrolled parallel corpora. However, a minority of projects (**?**; **?**; **?**) have addressed automated learning of translation rules from a controlled corpus of carefully elicited sentences.

Learning from controlled elicitation, like any method of automatic learning, is useful for rare languages that are not spoken by any computational linguists[1]. Native speaker informants provide data as required by the learning algorithm, but do not need to have technical knowledge. The resulting systems are automatically learned, but also consist of human-readable rules that can be extended and modified. In this paper we will review the challenges of automated rule learning from controlled corpora in the context of our AVENUE machine translation project.

While some of the issues we encounter are specific to our project, others are common to all systems that automatically elicit a controlled corpus. Other work in this area includes **?**, **?**, and **?**, which describe the construction of a controlled corpus based on linguistic research across languages. They compiled a list of linguistic features (such as number) together with the possible values these features take on across languages (such as singular, dual, plural, and paucal). A bilingual speaker then provides the system with necessary information regarding these features, namely what values a certain feature

[1] Controlled elicitation systems can also be used for major languages in order to learn an MT system from a small amount of data.

can take in the language in question. The elicited information is then used to build a machine translation system between the elicitation language and the elicited language. This work ties in well with our project as it relies on a bilingual speaker to compile a knowledge base for the system. However, in our system, the native speaker is only required to translate sentences and phrases and is not required to analyze the syntactic features and values.

? also use an expert user who translates well-chosen sentences and annotates the translations with glosses. The glosses are used to grow a knowledge base for a machine translation system. In order to build a machine translation system for a new language pair, expert users collect a set of representative sentences from such sources as grammar books. In the AVENUE project we also draw from such sources as well as from guides for field linguists **?**; **?**.

Although our learning approach can be applied to controlled or uncontrolled corpora, we have designed it specifically for controlled corpora. A controlled corpus can systematically target specific grammatical features and constructions that would be encountered sparsely and randomly in an uncontrolled corpus. Sparseness is especially problematic when the uncontrolled corpus is small, as it the case for the minor languages we are working with (for example, Mapudungun, the language of the Mapuche people of Chile). We believe that the collection of a controlled corpus may be a more time-efficient undertaking in the long run than collecting a large enough uncontrolled corpus. In the future, we will aim at incorporating data from both controlled and uncontrolled corpora.

2 Project Overview

In the development of AVENUE, we are targeting both language portability and shortened development time. The system architecture (Figure 1) is divided into a learning module and a run-time module. Within the learning module is an elicitation process for acquiring data and a machine learning module (divided into seed generation and seeded version space learning) for learning rules from the data. The rules are then used by the run-time system.

The focus of this paper is the elicitation module. Figure 2 shows the elicitation interface. A bilingual informant is prompted to translate a number of sentences and specify the alignment between source language words and target language words. The list of sentences, which we call the elicitation corpus, is designed to cover major linguistic phenomena in typologically diverse languages. The source language is a major language such as English or Spanish. The target language can in principle be any language. Currently we are working on Mapudungun (spoken in Chile) and are negotiating partnerships with other indigenous groups.

The second module is a system that uses the elicited sentences to automatically infer transfer rules. A process called seed generation predicts approximations to the desired transfer rules. Then a version space algorithm (based loosely on **?**) adjusts the predicted transfer rules to the correct level of generalization and weeds out faulty transfer rules. This second module is currently under construction. Table 1 shows the format of a transfer rule that is learned by the system.

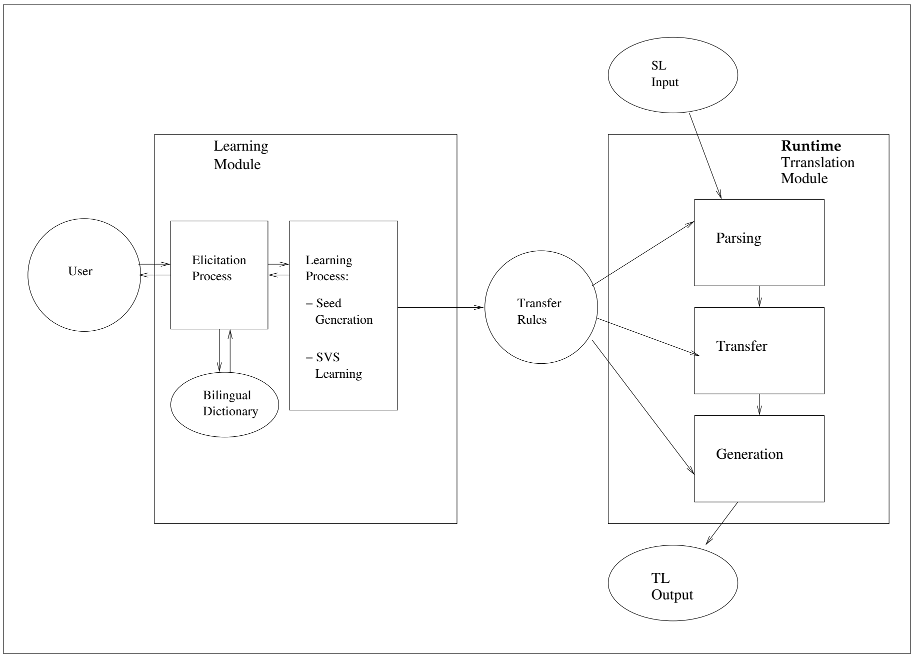

Figure 1: AVENUE system architecture

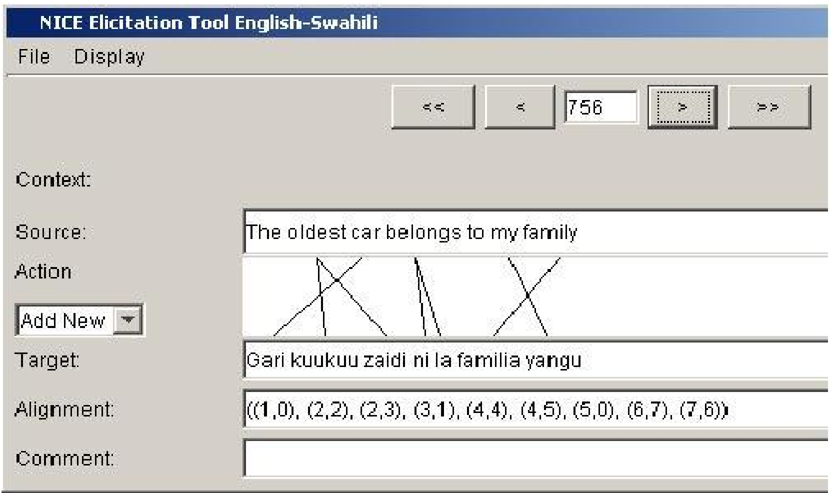

Figure 2: AVENUE elicitation interface

The elicitation corpus is a list of sentences in a major language like English or Spanish. The current pilot elicitation corpus has about 850 sentences, though we expect it to grow to at least the number of sentences in **?**, which includes around 6,000 to 10,000 sentences, and to ultimately cover most of the phenomena in the Comrie and Smith **?** checklist for descriptive grammars. In order to minimize the number of sentences an informant actually has to translate, we prune the corpus based on already

```
;; Hebrew Transfer Rule Example
English: the big boy
Hebrew: ha yeled ha gadol
NP::NP : [DET ADJ N] -> [DET N DET ADJ]
(;;X-Y Alignment
(X1::Y1)
(X1::Y3)
(X2::Y4)
(X3::Y2)
;;X-side constraints
((X1 NUMBER) = (X3 NUMBER))
((X1 DEFINITENESS) = +)
;;Y-side constraints
((Y2 NUMBER) = (Y4 NUMBER))
((Y2 GENDER) = (Y4 GENDER))
;;X-Y constraints
((X0 NUMBER) = (Y0 NUMBER))
((X0 DEFINITENESS) = (Y0 DEFINITENESS)))
```

Table 1: Sample Transfer Rule

elicited knowledge. For more details, see **?**.

The elicitation corpus has three organizational properties: the sentences are grouped into minimal pairs; the order of elicitation is compositional (starting with smaller phrases and combining them into larger phrases); and the content of the corpus is dynamic in the sense that the elicitation takes a different path depending on what has been found so far. In this paper we will focus on the organization of the corpus into minimal pairs.

A minimal pair can be defined as two sentences that differ only in one feature. Each sentence can be associated with a feature vector. Consider, for instance, the following three sentences:

(i) *The man saw the young girl.*

(ii) *The men saw the young girl.*

(iii) *The woman saw the young girl.*

The first and the second sentence differ merely in one feature, namely the number of the subject. Thus, they represent a minimal pair. However, sentences one and three also form minimal pair. They differ only in the gender of the subject.

The organization of the corpus into minimal pairs facilitates *feature detection*, an important component of our rule learning mechanism. During feature detection, we attempt to identify which grammatical phenomena occur in a language and which do not. The phenomena, or features, that we look for include, for example, agreement of subjects and verbs for number, person, and gender; marking of verbs for past time in contrast to present time; marking of verbs for past time in contrast to non-past time, and so on.

Minimal pairs support feature detection in the following way. To find out whether target language verbs inflect depending on the number of the subject we compare the target language translations (supplied by the bilingual speaker) of sentences 1 and 2 above. If the verbs are identical, we conclude that there is no subject-verb agreement. If the verbs are not identical, we assume that the difference is caused by the number of the subject. However, we must be careful about our conjectures. For example, if the two verbs do not differ, then this could be an indication that target language verbs do not agree with their subjects in number. However, it could also mean that the verb used for comparison represents an exception to a more common rule. We approach this problem in two ways. First, we never rule out the existence of a feature based on only one example (i.e., one minimal pair). Second, the results from feature detection are interpreted not as absolutes, but as tendencies, which guide our search through the space of possible transfer rules and give more preference to those rules that use features that were detected over those that use features that were not detected. However, the latter type of rule is not considered impossible in our algorithm, accounting for the imperfect results of feature detection. A more detailed discussion of our feature detection module can be found in **?**.

3 Challenge Number 1: The bilingual informant

Before we can design a controlled corpus, we need to define clearly what the characteristics of a typical user of the system would be. Of high priority in our work is not to expect the bilingual informants to know and understand linguistic terminology. However, we do hope, somewhat unrealistically, for logical and consistent data. The capabilities of an ideal user are as follows:

1. **Translate a minimal pair as a minimal pair (where possible):** For example, a Hebrew speaker translating possessive phrases has the option of expressing a possessor as a separate word or as a suffix. Faced with the minimal pair "my brother" and "my sister" (differing in the gender of the possessee), we would hope that the informant would use suffixes for possessors in both phrases (*axi* (brother-my) and *axoti* (sister-my)) or separate possessors in both cases (*ha-ax sheli* (the-brother of-me) and *ha-axot sheli* (the-sister of-me)). If the informant happens to mix the two constructions (e.g., *ha-ax sheli* (the-brother of-me) and *axoti* (sister-my)) our feature detection module may temporarily form the hypothesis that the form of the possessive construction differs for masculine and feminine possessees.[2]

   Translating a minimal pair as a minimal pair may, however, not be appropriate, for example, if there really is no corresponding minimal pair in the target language or if using a minimal pair forces the target language informant to unnaturally choose source-language-like structures. (See Challenge 8 below.)

2. **Be consistent about word boundaries and word alignments:** Many of the languages we plan to address do not have standard orthographic conventions. For example, the Mapudungun locative *mew* can be written as a suffix or as a separate

[2]The option of using a suffix is not available for all nouns, but is available for kinship terms. It would, therefore, be correct for the feature detection module to hypothesize that the form of the possessive construction differs for alienable and inalienable possession, or differs based on individual lexical items.

word.[3] We hope that users will segment and align words consistently, although we understand that for many languages, especially those written without spaces between words (such as Chinese), few speakers will be unerringly consistent in segmenting sentences into words.

3. **Be able to supply grammaticality judgements with minimal context:** The internal machine learning program that analyzes the elicited data strives to acquire transfer rules automatically. In general, machine learning programs yield better performance if they have the option of learning actively, "asking questions". In our system, "asking questions" amounts to the system proposing a translation using a transfer rule that is not completely confirmed. The translation is then presented to the user for a grammaticality judgment. If the user can act as an oracle for the system, and, even more ideally, correct the translation to the closest possible fit, automated learning will be much facilitated. Unfortunately, it may be difficult for native speakers to give appropriate grammaticality judgements unless they are following the logic of the elicitation process.

Because languages always include inconsistent options and unclear categories, we have designed our learning algorithm to learn from "messy" data. However, "clean" data will result in more general and correct rules. The above is, of course, a very idealized picture of teh bingual informant. While we cannot expect that an informant will fit this description perfectly (and make no mistakes), we have had positive experiences the our Mapuche informants in Chile.

4 Challenge Number 2: Morphology and the lexicon

We would like automated elicitation to proceed compositionally: starting with single words, proceeding to phrases, and then to clauses. (See Challenge 4.) As the starting point, lexicons or word lists may exist as a product of prior work on the language. However, if a word list or lexicon does not exist or is not complete, acquisition of lexical items and morphology poses a snag. Eliciting the entire vocabulary of a language one word at a time is tedious for the informant, and will result in only citation forms (e.g., nominative or infinitive). Eliciting words as part of sentences saves time, but results only in inflected forms. (Of course, one of the inflected forms might be the citation form.)

We prefer to elicit words as part of phrases or sentences as much as possible. Using the user-specified alignments, the learning system stores the translation of each word when it appears in the context of a sentence. This often results in multiple translations for one word, where the translations are morphological variants of each other. We are in the process of developing a morphology learning module that, combined with feature detection, can assign meanings to the morphological variants, and possibly identify word stems.

[3] Our partners in at Universidad de la Frontera in Chile happen to be linguists and are specifying a standard orthography for our project, but we cannot always count on partners having lingusitic expertise.

There are situations in which we may need to resort to extensive lexical acquisition separate from sentence elicitation. Word classes may have only partially predictable criteria for membership. For example, gender can be partially determined by phonological form in Spanish, noun class can be partially determined semantically in Bantu languages, and numerical classifiers can be partially determined by shape and meaning in Japanese. Our automatic feature detection module will determine whether or not such classes exist in a language, but the membership of each class may need to be determined one word at a time if the information is not found in a pre-existing lexicon.

5 Challenge Number 3: Learning grammatical features

As described above, our automatic feature detection module identifies grammatical features and agreements that occur in a language. There are two main challenges associated with eliciting grammatical features. The first is to achieve reasonably large coverage. It is clear that the number of grammatical features across languages is very large. Not only do we need to compile a typologically complete list of features and their possible values across languages. We also need to construct sentences that elicit each of these features. Luckily this task only needs to be done once in the design of the elicitation corpus. We are growing the elicitation corpus based on research on typology and universals (e.g. **?**; **?**).

The second challenge concerning grammatical features is that the source language (so far, English or Spanish) may not exhibit a grammatical feature that exists in the target language. This may require circumlocutions in the source language elicitation sentences. For example, English does not exhibit dual number. However, if we want to know whether another language has dual, we present the following minimal pair of sentences:

(i) *Two men ran across the street.*

(ii) *Three men ran across the street.*

If the translation of *men* differs in the target language, we have an indication that the target language marks nouns for dual. Not all grammatical features are as straightforward to elicit. However, research has been conducted on how to elicit information from users without misleading them **?**. These elicitation methods use media such as pictures, videos, audio clues, etc.

6 Challenge Number 4: Compositional elicitation

Compositional elicitation is based on the principle that sentences are constructed from smaller units, such as noun phrases, adjective phrases, etc. When constructing a transfer rule for a sentence, it will be helpful to the system if transfer rules for these smaller units have already been learned. For instance, consider the following simple sentence:

The women	*danced*
NP	VP

If the system has already learned an NP transfer rule for *the women*, then the learning the transfer rule for the entire sentence is an easier task than otherwise. At first glance, compositional elicitation sounds like a time-efficient and effort-minimizing idea. If we can elicit translations for noun phrases, learn their transfer rules, then do the same for clauses, we can put the rules together to translate whole sentences.

However, as was hinted at above, compositional elicitation becomes more complex when dealing with a language that marks noun phrases in different functions in the sentence. For instance, German marks direct objects by using accusative case marking, but if a noun phrase is elicited in isolation, the translation would be in the nominative case. This puts an additional challenge on compositional learning. Not only do we need to learn transfer rules for phrases, we also need to learn in what context they appear in what specific form. The result is a kind of ordering paradox: We want to learn the smaller pieces first, but we cannot learn them without the larger context.

In AVENUE we have applied a simple bootstrapping loop to get around the paradox. We elicit the simplest noun phrases in the context of very short sentences. We use these short sentences to determine how noun phrases are affected by the sentence (e.g., case marking) and how they affect the rest of the sentence (e.g., subject-verb agreement, alternative constructions for inanimate subjects, etc.) We then elicit more complex sentences.

7 Challenge Number 5: Elicitation of non-compositional data

In certain cases, we can expect a priori that the source and target language sentences will differ widely in their grammatical construction. Many fixed expressions, for example expressions of greeting, polite conversation, etc., cannot easily be captured in syntactic transfer rules. In fact, in our other projects, we have capitalized on the inherent non-compositionality of some types of sentences, such as those found in task oriented dialogues **?**. The examples below show fixed expressions that often do not translate literally between two languages, here German and English.

(1)

Guten	*Tag.*
good.ACC.SG	day.ACC

'Hello.'

(2)

Wie	*geht*	*es*	*Ihnen?*
how	go.3SG.PRES	it	you.DAT

'How are you?'

(3)

Wieviel	*kostet*	*das?*
how-much	cost.3SG.PRES	this.NOM

'How much is this?'

(4)

Ich	*wünsche*	*Ihnen*	*einen*	*schönen*	*Tag.*
I.NOM	wish.1SG.PRES	you.DAT.POLITE	a.ACC.SG.M	nice.ACC.SG.M	day

'Have a nice day.'

(5)

Es	*ist*	*nett,*	*Sie*	*kennen zu lernen.*
it.NOM	be.3SG.PRES	nice	you.ACC.POLITE	meet.INF

'It is nice to meet you.'

Meanings that are frequently expressed non-compositionally will be elicited in separate sections of the corpus. These will include greetings, introductions, polite request forms (*Would it be possible for you to ...*), and so on. We do not try to learn transfer rules for subcomponents of these sentences, nor do we try to apply previously learned (phrase) transfer rules to these sentences. In the future, however, we will address the problem of exploiting the small degree of compositionality still present in these problem cases. For example, we would like to have a transfer rule that covers both *How are you?* and *How is your brother?*.

8 Challenge Number 6: Verb subcategorization

We have identified the elicitation of clauses as a particular challenge. The main reason for this is that verbs do not agree in their subcategorizations across languages (as discussed in the literature on machine translation divergences, e.g. **?**, p. 124). For example, in the English construction *He declared him Prime Minister*, the verb *declare* subcategorizes for two noun phrases, an indirect object (*him*), and a direct object (*Prime Minister*). The same construction in German would be:

(6)	*Er*	*ernannte*	*ihn*	*zum*	*Premierminister.*
	he	declare.3SG.PAST	him.ACC	to+the.DAT.SG	prime-minister

‘He declared him prime minister.’

It can be seen that in German the verb "ernennen" ("to declare") subcategorizes for direct object ("ihn" - "him"), as well as a prepositional phrase introduced by the preposition "zu" ("to"). Subcategorization mismatches are common, and may be random or systematic according to verb classes (see **?**). Thus our aim will be lexical transfer rules for individual verbs or verb classes as opposed to lexically independent clause rules.

9 Challenge Number 7: Alignment issues

As has been pointed out in the literature (e.g. **?**), human-specified alignment is not noise-free. There is some degree of disagreement between people who align the same sentences, and even one and the same person will sometimes not be completely consistent. This provides a problem for a system that automatically elicits data and relies heavily on informant-specified alignments. Informants can be given intuitive instructions for alignment - for instance, we will ask the informant to align a word to only those words in the target language that actually have the root with the same meaning. Such guidelines have proven useful, as in **?**. Yet, the learning process must still be tolerant to noise in alignments, and we must also be prepared to accept non-optimal rules that are learned from noisy data.

10 Challenge Number 8: Bias toward the source language

A pervasive problem with elicitation is that the response might be biased by the stimulus. This is particularly problematic when the elicitation language (e.g., English) has fixed word order and the target language has free word order. In many languages the order of words actually reflects discourse context such as old and new information. Bias toward an English-like word order would create problems of text coherence and lose critical information. Multi-modal, context-based elicitation, however, would be prohibitively slow. Developing methodology for efficient context-based elicitation is therefore an ongoing concern.

11 Conclusions and Future Work

We have presented challenges that we face when designing and implementing a tool that elicits a controlled bilingual corpus from an informant. Some of these challenges are common to all systems that elicit linguistic data, for example alignment issues and acquisition of morphemes. Others are unique to our system because they are related to the way we use the elicited data. In particular, the elicited data is largely processed automatically. This poses special problems when dealing with noisy data, either because a language behaves differently than the designers anticipated, or because the informant makes mistakes. In our rule learning system, we have to have explicit noise-handling procedures, because we cannot expect to receive noise-free data from the informant.

Our long-term plan is to expand the existing elicitation corpus to have reasonable coverage of linguistic phenomena that occur across languages and language families. In practice, we first have to focus on the more common features, so that the informant's time is used as efficiently as possible and a preliminary rough translation can be produced.

References

Bouquiaux, Luc & J.M.C. Thomas: 1992, *Studying and Describing Unwritten Languages*, Dallas, TX: The Summer Institute of Linguistics.

Comrie, Bernard: 1981, *Language Universals & Linguistic Typology*, Chigaco, IL: The University Chicago Press, 2nd edn.

Comrie, Bernard & N. Smith: 1977, 'Lingua descriptive series: Questionnaire', *Lingua*, **42**: 1–72.

Dorr, Bonnie: 1992, *Machine Translation: A View from the Lexicon*, MIT Press.

Eskenazi, Maxine: 1999, 'Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype', *Language Learning & Technology*, **2**(2): 62–76.

Greenberg, Joseph H.: 1966, *Universals of language*, Cambridge, MA: MIT Press, 2nd edn.

Jones, Douglas & R. Havrilla: 1998, 'Twisted pair grammar: Support for rapid development of machine translation for low density languages', in *AMTA*.

Levin, Lori, Donna Gates, Alon Lavie & Alex Waibel: 1998, 'An interlingua based on domain actions for machine translation of task-oriented dialogues', in *ICSLP*, vol. 4, pp. 1155–1158.

Melamed, Dan I.: 1998, 'Manual annotation of translational equivalence: The Blinker project', Tech. Rep. 98-07, IRCS.

Mitchell, Tom: 1982, 'Generalization as search', *Artificial Intelligence*, **18**: 203–226.

Nirenburg, Sergei: 1998, 'Project boas: A linguist in the box as a multi-purpose language', in *LREC*.

Nirenburg, Sergei & V. Raskin: 1998, 'Universal grammar and lexis for quick ramp-up of MT systems', in *COLING-ACL*.

Probst, Katharina, Ralf Brown, Jaime Carbonell, Alon Lavie, Lori Levin & Erik Peterson: 2001, 'Design and implementation of controlled elicitation for machine translation of low-density languages', in *Workshop MT2010 at Machine Translation Summit VIII*.

Sherematyeva, Svetlana & Sergei Nirenburg: 2000, 'Towards a unversal tool for NLP resource acquisition', in *LREC*.

Trujillo, Arturo: 1999, *Translation Engines*, Springer.