Using Corpora to Develop Limited-Domain Speech Translation Systems

Manny Rayner
SRI International
Suite 23, Millers Yard
Cambridge CB2 IRQ
United Kingdom
many@cam.sri.com

Pierrette Bouillon
ISSCO
University of Geneva
54, route des Acacias
1227 Geneva, Switzerland
pb@divsun.unige.ch

David Carter
SRI International
Suite 23, Millers Yard
Cambridge CB2 IRQ
United Kingdom
dmc@cam.sri.com

Abstract

The paper describes the Spoken Language Translator (SLT) system, a prototype automatic speech translator. SLT is currently capable of translating spoken English queries in the domain of air travel planning into either Swedish or French, using a vocabulary of about 1200 words. We present an overview of the system's architecture, concentrating on how rationally constructed balanced corpora are used to allow rapid development of high-quality limited-domain translation systems.

1 Introduction

During the last five years, people have started to believe there is a serious possibility of building practically useful spoken language translators for limited domains. There are now a number of high-profile projects with large budgets, the most well-known being the German Verbmobil effort. At the moment, the best systems are at the level of advanced prototypes; making projections from current performance, it seems reasonable to hope that these could be developed into commercially interesting systems within a time-scale of about another five years.

This paper describes a project centered around one such advanced prototype, the Spoken Language Translator (SLT) system. SLT can translate spoken English utterances from the domain of air travel planning (ATIS; (Hemphill et al., 1990)) into spoken Swedish or French, using a vocabulary of about 1200 stem entries. Table 1 shows some examples of typical sentences from this domain, together with French translations produced by the system. The Swedish version of SLT has been operational since June 1993, and has been publicly demonstrated on numerous occasions. The French version became operational fairly recently, and was publicly demonstrated for the first time at the Language Engineering Convention in London in October 1995. An initial French-to-Spanish version of the system also exists. By the middle of 1996, we expect to have developed SLT further to permit translation of spoken Swedish and French into English, allowing translation of two-way dialogue. The system is configured from general-purpose speech and language processing components, making it possible to reconfigure it fairly easily to new language pairs and application areas.

The rest of the paper is organised as follows. In Section 2, we describe the SLT system at a coarse level of detail, concentrating on issues specifically related to translation. Section 3 constitutes the main content of the paper, describes in more detail how we use rationally constructed balanced corpora to allow rapid development of limited-domain translation systems. Section 4 concludes.

2 Overview of SLT

The SLT system and its main components are documented in great detail in numerous technical reports; Section 5 below describes how to get hold of them. Here, we will attempt to describe the system's main components in a non-technical way, from the point of view of a person interested in knowing how they relate to the translation-specific problems that are the main focus of the paper.

The SLT system consists of a set of individual processing modules, linked together in a pipelined fashion as shown in Figure 2. Speech enters the system at the top left of the diagram and is processed by the SRI DECIPHER(TM) recognizer (Murveit et al., 1993). DECEPHER(TM) is an advanced

1	English	What is the earliest flight from Boston to Atlanta?
	French	Quel est le premier vol Boston-Atlanta?
2	English	Show me the round trip tickets from Baltimore to San Diego
	French	Indiquez-moi les billets aller-retour Baltimore-San Diego!
3	English	I would like to go about nine a.m.
	French	J'aimerais partir aux environs de neuf heures.
4	English	Show me the fares for Continental flight one forty seven
	French	Indiquez-moi les tarifs du vol Continental cent quarante-sept!
5	English	Show me ground transportation for Dallas
	(heard)	Show me ground transportation from Dallas
	French	Indiquez-moi les transports en partance de Dallas!
6	English	Which of these flights are nonstop?
	French	Lesquels de ces vols sont directs?
7	English	What flights are available from Pittsburgh to Memphis on United?
	French	Quels vols y a-t-il de Pittsburgh à Memphis avec United?
8	English	How many cities are served by Delta with first class flights?
	French	Combien de villes dessert Delta avec des vols en première classe?
9	English	Are there any flights from Boston to San Francisco which stop in Denver?
	French	Y a-t-il des vols Boston-San Francisco qui font escale à Denver?
10	English	I would like to fly from Tacoma to New York with a stopover in Washington.
	French	J'aimerais aller de Tacoma à New York avec escale à Washington.
11	English	What flights are available on Wednesday afternoon from Charlotte to Los Angeles?
	French	Quels vols y a-t-il le mercredi de Charlotte à Los Angeles après-midi?
12	English	What are the arrival times in Washington D C?
	(heard)	What are all the arrival times in Washington D C?
	French	Quelles sont toutes les heures d'arrivée à Washington?
13	English	Does Northwest flight two oh two serve meals?
	French	Est-ce que le vol de Northwest deux cent deux sert les repas?
14	English	How long does it take to fly from Milwaukee to Boston?
	French	Combien de temps faut-il pour aller de Milwaukee à Boston?

Table 1: SLT translation examples. The English sentences were randomly selected from the syntactic representative corpus and read out once each by one of the authors (a British English native speaker) using a close-talking microphone in a quiet office environment. In examples 5 and 12, the utterance was misrecognized by the system; the middle line shows what it believed it heard. The level of performance can be regarded as typical for a user who knows the domain and is familiar with using the system.

speaker-independent continuous speech system, which for tasks like ATIS has a word error rate of less than 5%. The recognizer produces an N-best list of hypotheses, and passes them to the source language processor, a copy of the SRI Core Language Engine (CLE; (Alshawi (ed), 1992)) loaded with an English grammar and lexicon. The CLE produces for each speech hypothesis a set of possible analyses in a logic-like representation called Quasi Logical Form (QLF). Once the set of possible QLF analyses has been collected, the CLE uses a trainable preference method to select the most plausible one, (Alshawi and Carter, 1994; Rayner et al., 1994a). The recognizer is accurate enough that speech/language techniques such as word-spotting and robust recognition are not needed; good performance is achieved by the comparatively simple methods described above.

The QLF analysis selected as most plausible is passed to the transfer component, which first annotates it with extra information in a rule-based pre-transfer phase. This phase is mainly used to handle simple cases of metonymy. Next, a set of possible target-language QLFs is created, using a set of unification-based transfer rules (Alshawi et al., 1991). The target QLFs are stored in a "packed" form (Tomita, 1986; Rayner and Bouillon, 1995) to avoid a combinatoric explosion when many transfer choices are non-deterministic. Following this, a second set of trained statistical preferences extract the most plausible transferred QLF and "unpack" it into a normal representation. A rule-based post-transfer phase then performs some simple rewriting of the transferred QLFs; the most important operation is reordering of modifiers.

When the "best" target-language QLF has been selected, unpacked, and subjected to the post-transfer operations, it is passed to a second copy of the CLE, loaded with a target-language grammar and lexicon; this generates a surface string using the Semantic Head-Driven algorithm (Shieber et al., 1990). Finally, the target-language string is passed to a speech synthesizer and converted into output speech. Speech synthesis is handled by the Telia Prophon system for Swedish, and by the CNETVOX text-to-speech system for French.

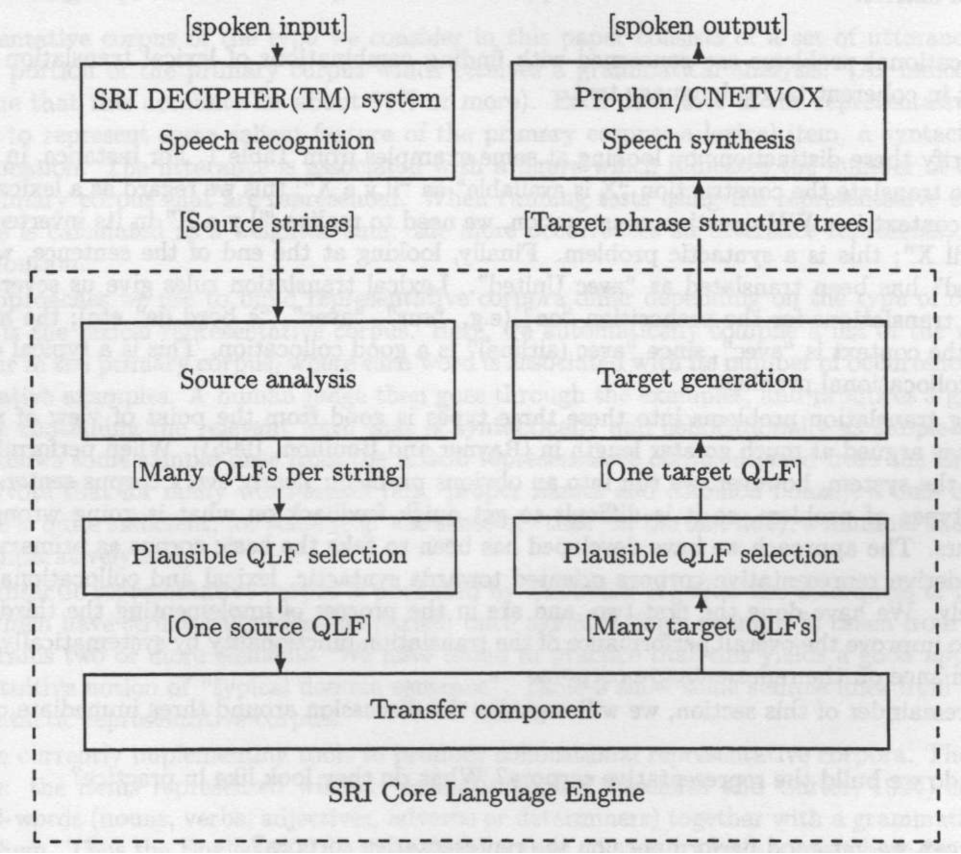

Figure 1: Basic SLT processing

3 Corpus based development

In this section, we will explain in more detail how we develop limited-domain translation applications within the SLT framework. Our basic approach is empirical and corpus-driven. We start by collecting a substantial source-language corpus, typically on the order of 5000 utterances or more. For the ATIS domain (the one we have devoted most time to), the English corpus was created by letting naive users interact with real and simulated natural language interfaces to airline flight information databases; versions of the corpus in other languages were then produced by translating the original corpus. The translation effort was divided between a large number of different translators in order to ensure a suitable diversity in style and word-choice.

When the corpus has been created, the next step is to do the work necessary to achieve good grammar and lexicon coverage on the source-language side, so that at least 85-90% of all utterances are assigned a correct syntactic/semantic analysis. The problems involved here are mainly of two types: extension of the lexicon with suitable domain-specific entries, and training of the preference mechanisms that distinguish plausible from implausible grammatical analyses. These issues are discussed at length elsewhere (cf. in particular Chapters 5 and 7 of (Agnäs et al., 1994)). In the present paper, we will assume that the work needed to achieve a high level of source-language coverage has already been carried out, and concentrate on the problems that arise in the translation step.

We find it helpful to divide translation problems into three different types: *syntactic, lexical* and *collocational.* We use these terms in the following (slightly non-standard) way.

- Syntactic problems are concerned with finding a translation that has the right syntactic structure, in terms of using the right grammar rules in the right configuration.

- Lexical problems are concerned with finding locally correct choices for translation of individual lexical entries.

- Collocational problems are concerned with finding combinations of lexical translation rules that result in coherent target language text.

Let us clarify these distinctions by looking at some examples from Table 1. For instance, in example 7 we want to translate the construction "X is available" as "il y a X"; this we regard as a lexical problem. Since the context is a WH-question construction, we need to realize "il y a X" in its inverted form, i.e. as "y a-t-il X"; this is a syntactic problem. Finally, looking at the end of the sentence, we see that "on United" has been translated as "avec United". Lexical translation rules give us several possible candidate translations for the preposition "on" (e.g. "sur", "avec", "à bord de" etc); the appropriate choice in the context is "avec", since "avec (airline)" is a good collocation. This is a typical example of a simple collocational problem.

Dividing translation problems into these three types is good from the point of view of modularity (the point is argued at much greater length in (Rayner and Bouillon, 1995)). When performing routine testing of the system, however, we run into an obvious problem: nearly every corpus sentence exhibits all three types of problem, so it is difficult to get quick feedback on what is going wrong when an error occurs. The approach we have developed has been to take the basic corpus as primary input; we use it to derive representative corpora oriented towards syntactic, lexical and collocational problems respectively. We have done the first two, and are in the process of implementing the third. We then attempt to improve the overall performance of the translation functionality by systematically improving its performance on the representative corpora.

In the remainder of this section, we will organize the discussion around three immediate questions:

- How do we build the representative corpora? What do they look like in practice?

- How can we get good performance on the representative corpora?

- If we get good performance on the representative corpora, will we then get good performance on real unseen data as well?

Freq.	Word	Example
1915	to *(prep.)*	to Boston.
1842	from	from Boston.
1478	flights	the flights.
1086	flight	a flight.
1047	the *(sing.)*	the flight.
900	on	on Wednesday.
805	me	show me all flights.
776	is *(copula)*	what is the flight?
768	what *(sing.)*	what is the flight?
756	Boston	Boston.
…	…	…
155	that *(rel. pro.)*	a flight that stops in Boston.
…	….	…
39	is *(aux. v.)*	what aircraft is used on this flight?
…	…	…
31	that *(demon.)*	that.
…	…	…
6	take *(t. v.)*	what flight can I take.
5	take *(extr. v.)*	how long does it take to get to Boston?

Table 2: Sample lines from English ATIS lexical representative corpus.

3.1 Building representative corpora for MT applications

A representative corpus of the type we consider in this paper consists of a set of utterances, derived from the portion of the primary corpus which receives a grammatical analysis. (As indicated above, we assume that this accounts for about 90% or more). Each utterance in the representative corpus is intended to represent some salient feature of the primary corpus; a lexical item, a syntactic pattern, or a collocation. The utterance is associated with a figure which indicates the number of occurrences in the primary corpus that are represented. When running tests using the representative corpora, the test score is calculated as a weighted sum: the more occurrences an utterance represents, the greater its contribution.

The approaches we use to build representative corpora differ depending on the type of corpus. The simplest is the lexical representative corpus. Here, we automatically compile a list of the word-senses that occur in the primary corpus, where each word is associated with its number of occurrences and a set of illustrative examples. A human judge then goes through the examples, and produces a grammatical utterance containing the relevant word that is syntactically and collocationally as simple as possible. Table 2 shows some sample lines from the lexical representative corpus derived from the English ATIS corpus. Note that for many word-senses (e.g. proper names and common nouns), a one- or two-word utterance is quite sufficient; for some (e.g. the sense of "take" in the last line), a minimal example needs to be comparatively long.

The syntactic representative corpus is produced by clustering together the utterances in the primary corpus which have structurally identical parses: once again, a single example is taken from each group that contains two or more elements. We have found in practice that this yields a good approximation to the intuitive notion of "typical domain sentence". Table 3 show some sample lines from the English ATIS syntactic representative corpus.

We are currently implementing tools to produce collocational representative corpora. The basic idea is simple: the items represented will be "semantic triples" (Alshawi and Carter, 1994) consisting of two head-words (nouns, verbs, adjectives, adverbs or determiners) together with a grammatical relation linking them. Thus the English ATIS collocational corpus will for example contain high-frequency lines for items like "flight to (airport)", "flight from (airport)", "show direct_obj flight", "fly with (airline)", "bare_singular det_of transportation" and so on. Software for automatically extracting these triples already exists, and only needs a fairly minimal amount of adaptation to be used for the purposes

Freq.	Pattern	Example
76	v np det nbar p np p np	Show me all flights from Pittsburgh to Atlanta.
60	advp	One way.
45	np v det adjp nbar p np p np	What is the first flight from Boston to Dallas?
39	v np nbar p np p np	Show me flights from Atlanta to Baltimore.
33	np v det adjp adjp nbar p np p np	What is the cheapest one way fare from Boston to Dallas?
31	p np	To Denver.
31	np v det nbar p np p np	What are the flights from Boston to Atlanta?
…	…	…
14	np v v v p np p np	I want to fly from Atlanta to Philadelphia.
14	np p np np	Philadelphia to Boston Monday.
14	det nbar nbar v adjp p np	What ground transportation is available in Boston?
5	v np det nbar v p det nbar	Is there a meal served on that flight?
5	v np det nbar p np p np np v p np	Are there any flights from Boston to San Francisco which stop in Denver?

Table 3: Sample lines from English ATIS lexical representative corpus.

envisaged here.

3.2 How well can we do on representative corpora?

We have carried out extensive testing on syntactic and lexical representative corpora in several language pairs. Our overall conclusion is that it is perfectly realistic to aim for frequency-weighted scores of around 99-99.5% on lexical corpora, and 95-97% on syntactic corpora. Scores like these can be attained by an effort on the order of 4-8 person-months for domains of the kind of complexity we have considered in SLT.

Most of the remaining errors are due to one of two causes. Firstly, source-language analysis can of course fail to produce a valid result. This is an important problem, but one which is outside the scope of the current paper: it is discussed in detail in Chapters 5 and 7 of (Agnäs et al., 1994). For instance, sentence 11 in Table 1 is an error of this type: the root cause of the bad translation is that "Wednesday afternoon from Charlotte to Los Angeles" has incorrectly been parsed as a noun-phrase.

The second common type of error is the collocational error: translation is syntactically correct, but the word-choice is wrong due to insufficiently exact modelling of the collocational properties of the target language. Sentence 13 in Table 1 is an example of a collocational error; the implicit bare plural determiner of "meals" has been translated as "les" (the default), instead of "des", which is the correct choice in this context. We have made fair progress towards dealing with collocational problems (Rayner and Bouillon, 1995), but still need to do more work. We hope that production of a collocational representative corpus will make it simpler to isolate and solve collocational problems in the same systematic way in which we currently attack syntactic and lexical ones.

3.3 Are representative corpora useful?

Representative corpus scores on the level described above correspond to scores around 75-85% for real unseen utterances of lengths manageable by current state-of-the-art speech recognizers. (In practice, we have applied a limit of 15 words). This may not sound particularly high. We have discovered, however, that since nearly all commonly occurring constructions are covered, there is usually some obvious rephrasing of a failing utterance which passes translation: users can use the intuitively correct strategy of repeating what they said in a simpler way. This gives rise to a system that feels more or less habitable.

4 Conclusions

It is currently fashionable to say that the notion of "representative corpus" is impossible to define in a meaningful way. We agree that care needs to be taken, but our experiences persuade us that the type of representative corpus described here admits of a rigorous definition and is also of considerable practical utility. We hope that this paper will encourage other people to re-examine the idea and develop it further.

5 More information about SLT

Most of the technical reports describing SLT are available on-line from SRI International; for the Internet user, the simplest way to get hold of these is by accessing the WWW URL

http://www.cam.sri.com

In particular, the SLT first-year report (Report CRC-043; pp 159) gives a detailed picture of the system as of early 1994.

The Core Language Engine (editor, Hiyan Alshawi; MIT Press, 1992; pp 322) presents an in-depth account of the basic CLE system.

Information about the DECIPHER(TM) recognizer is available from the SRI Speech Technology and Research Laboratory Internet users are advised to access WWW URL

http://www-speech.sri.com

6 Acknowledgements

Work on the generic and Swedish-specific aspects of the Spoken Language Translator project is funded by Telia Research AB, Stockholm, Sweden. French-specific work is funded by SRI International and Suissetra.

Apart from ourselves, the following people have contributed to the development of SLT: Marie-Susanne Agnäs, Hiyan Alshawi, Anita Andersson, Ivan Bretan, George Chen, Ken Cedar, Michael Collins, Dick Crouch, Vassilios Digalakis, Barbro Ekholm, Robert Eklund, Mikael Eriksson, Gabriela Fernandez, Björn Gambäck, Shih-Chi Huang, Jaan Kaja, Jussi Karlgren, Martin Keegan, Bertil Lyberg, Leo Neumeyer, Dominique Petitpierre, Patti Price, Steve Pulman, Christer Samuelsson, Tomas Svensson and Mats Wirén.

References

Agnäs, M-S., Alshawi, H., Bretan, I., Carter, D.M. Ceder, K., Collins, M., Crouch, R., Digalakis, V., Ekholm, B., Gambäck, B., Kaja, J., Karlgren, J., Lyberg, B., Price, P., Pulman, S., Rayner, M., Samuelsson, C. and Svensson, T. 1994. Spoken Language Translator: First Year Report. SRI technical report CRC-043 (also SICS research report R94:03) Available through WWW from http://www.cam.sri.com

Alshawi, H., Carter, D., Rayner, M. and Gambäck, B. 1991. Transfer through Quasi Logical Form. Proc. 29th ACL, Berkeley, CA.

Alshawi, H. (ed.) 1992. The Core Language Engine. MIT Press.

Alshawi, Hiyan, and David Carter. 1994. Training and Scaling Preference Functions for Disambiguation. Computational Linguistics, 20:4.

Hemphill, C.T., J.J. Godfrey and G.R. Doddington. 1990. The ATIS Spoken Language Systems pilot corpus. Proc. DARPA Speech and Natural Language Workshop, Hidden Valley, Pa.

Murveit, H., Butzberger, J., Digalakis, V. and Weintraub, M. 1993. Large Vocabulary Dictation using SRI's DECIPHER(TM) Speech Recognition System: Progressive Search Techniques. Proc. Inter. Conf. on Acoust., Speech and Signal, Minneapolis, Minnesota.

Rayner, M., Alshawi, H., Bretan, I., Carter, D.M., Digalakis, V., Gambäck, B., Kaja, J., Karlgren, J., Lyberg, B., Price, P., Pulman, S. and Samuelsson, C. 1993. A Speech to Speech Translation System Built From Standard Components. Proc. 1st ARPA workshop on Human Language Technology.

Rayner, M., D. Carter, V. Digalakis and P. Price. 1994. Combining Knowledge Sources to Reorder N-Best Speech Hypothesis Lists. Proc. 2nd ARPA workshop on Human Language Technology.

Rayner, M. and P. Bouillon. 1995. Hybrid Transfer in an English-French Spoken Language Translator. Proc. of IA '95, Montpellier, France.

Shieber, S. M., van Noord, G., Pereira, F.C.N and Moore, R.C. 1990. Semantic-Head-Driven Generation. Computational Linguistics, 16:30-43.

Tomita, M. 1986. Efficient Parsing for Natural Language. Kluwer Academic Publisher.