Representational and architectural issues in a limited-domain medical speech translator

Manny Rayner (1), Pierrette Bouillon (1), Marianne Santaholma (1),
Yukie Nakao (2)
(1) University of Geneva, TIM/ISSCO
40, bvd du Pont-d'Arve,
CH-1211 Geneva 4, Switzerland
mrayner@riacs.edu, Pierrette.Bouillon@issco.unige.ch,
Marianne.Santaholma@eti.unige.ch
(2) National Institute for Communications Technology,
3-5 Hikaridai, Seika-cho, Soraku-gun,
Kyoto, Japan 619-0289
yukie-n@khn.nict.go.jp

Mots-clefs : reconnaissance de la parole, traduction de la parole, aide au diagnostic médical

Keywords: speech understanding, speech translation, computer-aided diagnosis

Résumé Cet article dresse un aperçu du système MedSLT, un système de traduction de la parole dans le domaine médical pour un vocabulaire limité. Il met l'accent sur le problème du choix du type de représentation pour les constructions temporelles et causales. Nous montrons que celles-ci ne peuvent pas être représentées par des structures plates, généralement utilisées pour ce type d'application, mais qu'elles nécessitent des stuctures plus riches, enchâssées, qui permettent d'obtenir une traduction plus adéquate. Nous expliquons comment produire ces représentations et écrire des règles de traduction économiques qui mettent en correspondance les représentations sources dans la représentation interlingue correspondante

Abstract We present an overview of MedSLT, a medium-vocabulary medical speech translation system, focussing on the representational issues that arise when translating temporal and causal concepts. Although flat key/value structures are strongly preferred as semantic representations in speech understanding systems, we argue that it is infeasible to handle the necessary range of concepts using only flat structures. By exploiting the specific nature of the task, we show that it is possible to implement a solution which only slightly extends the representational complexity of the semantic representation language, by permitting an optional single nested level representing a subordinate clause construct. We sketch our solutions to the key problems of producing minimally nested representations using phrase-spotting methods, and writing cleanly structured rule-sets that map temporal and phrasal representations into a canonical interlingual form.

1 Introduction

As a subject, automatic speech translation is now a little more than ten years old. First generation systems, like Verbmobil (Wahlster, 2000), Spoken Language Translator (Rayner *et al.*, 2000) and Janus III (Lavie *et al.*, 1997) were essentially proofs of concept. We are now progressing to the stage where people want to build systems that have some claim to be useful: prominent recent examples are NESPOLE! (Lavie *et al.*, 2001), Tongues (Black *et al.*, 2002) and Phraselator (Phraselator, 2004). For obvious reasons, one application area that stands out is medical translation; this paper will focus on representational issues in MedSLT, a medium-vocabulary medical speech translation system (Rayner & Bouillon, 2002; Rayner *et al.*, 2003a).

There are many different contexts in which medical translation could potentially be useful. In the scenario targeted by the MedSLT system, we envisage that a doctor wishes to perform a preliminary examination of a patient who does not speak the doctor's language. The system allows the doctor to pose normal examination questions in her own language, translating them into the patient's language. The task appears tractable, given current speech technology. Medical examination questions are fairly stereotypical. It is also feasible for the dialogue to be one-way, with the patient responding non-verbally, so the user (i.e. the doctor) can reasonably be assumed to have had time to acclimatize themselves to the system and learn its capabilities.

The first question we need to ask is what metrics we should use to evaluate the success or failure of the system; evaluation of machine translation is notoriously difficult, and must normally be carried out with reference to a specific task. In the context of the MedSLT task, the translation system is basically a diagnostic tool; thus, the critical question is whether the patient's responses will give the doctor misleading information. This has several implications. There is no particular requirement that translations be completely literal; often, a non-literal translation will be as good, or indeed better. It is however important for translations to be concrete and clear in meaning, even if this involves losing nuances in the source utterances — this contrasts sharply with many translation and interpretation tasks, where nuances of meaning can be vital. Above all, the system must be extremely reliable, since the consequences of a mistranslation can be serious.

Putting these requirements together, we arrive at the basic design. Given the uncertainty inherent in current speech understanding and machine translation technology, processing cannot be fully automatic. We always need to make sure that the system has understood correctly before it translates. Since translation will not in general be completely literal, the system needs to echo back to the source-language user an accurate paraphrase of the translation it proposes to ask. The user will then have the option of either approving the translation and proceeding, or else aborting.

As always, there is tension between precision and recall. If linguistic coverage is too restricted, the system becomes hard to use. However, robust coverage must be balanced against the potentially very serious consequences of a mistranslation in a safety-critical task. Precision, which in this context is going to mean precision on the utterances which the user approved for translation, is thus more important than recall.

Here, we will be particularly concerned with the representations used by the MedSLT system for source, target and interlingual levels of structure. Following on from the previous points, the key issues are the following:

- If we want to prioritise reliability, we prefer to have a tightly constrained set of representational primitives.

- If the system is going to have adequate coverage, it needs to be able to represent all the relevant concepts in the domain. In this domain, the problematic cases mostly involve temporal and causal relations.

- With regard to the abstract structure of the representation, the critical dimension is the opposition between nested and flat representions. Nested representations (parse trees, logical forms etc) are more fine-grained, and make it easier to support a wide range of constructs. Conversely, flat representations hide linguistic structure, but are inherently more robust. In particular, they are well-suited to speech understanding architectures based on phrase-spotting and other methods suitable for processing noisy input. This point is sufficiently important that many researchers working in spoken language understanding simply take for granted that all semantic representations will be flat lists of key/value pairs; (Young, 2002) provides a good overview of current trends here. Flat structures also greatly simplify the task of writing translation rules which define correspondences across widely differing language pairs.

In the rest of the paper, we describe the representational solution we have developed for MedSLT. Linguistic representations are flat enough that it is simple to write phrase-spotting patterns and translation rules, but sufficiently expressive that they can capture all the key concepts of the domain. In the following sections, we first present the system; we then explain our approach focussing on the temporal and causal constructions. The central idea is to reduce causal concepts to temporal ones, which greatly simplifies the range of concepts that needs to be represented.

2 The MedSLT system

This section provides a brief overview of the current MedSLT prototype. The system is built on top of the Nuance toolkit platform (Nuance, 2003), and offers speech-to-speech translation from English into French, Japanese and Finnish[1]. It supports three separate medical diagnosis subdomains (headaches, chest pain, and abdominal pain) well enough that the full range of routine examination questions for each subdomain is covered. The vocabulary for each subdomain is between 300 and 450 words.

Translation is one-way in the doctor to patient direction, which means that most communication is in the form of yes-no questions that can be answered non-verbally. The system has a limited notion of dialogue context, so that it is possible to ask elliptical follow-on questions. For example, if the preceding question was "Is the pain sharp?", then "dull?" will be interpreted as "Is the pain dull?". Supporting ellipsis compensates to some extent for the restriction to yes-no questions. Instead of asking a single WH-question ("Where is the pain?", the doctor can ask an initial yes-no question with a series of elliptical follow-ups ("Is the pain in the front of the head?"... "The back of the head?"... "The left side?"... "The right side?")[2].

[1] Versions with French, Japanese and Spanish input and Spanish output are in various stages of preparation.

[2] The system does in fact also support WH-questions, since several doctors said they would like the option of using them as introductions to yes-no questions: "Where is the pain?"... "Is it in the front of the head?"

There are two versions of the system, using different speech understanding components. At the start of the project, we felt that there was a case to be made for using grammar-based recognition methods. Initially, we had no training data for creating statistical language models; also, the system is designed for expert users, and an earlier study we had been involved in (Knight *et al.*, 2001) suggested that grammar-based recognition can be more suitable for this type of user. These arguments are obviously not particularly strong. We wanted to be able to compare grammar-based speech understanding with a more standard architecture based on statistical language modelling and robust parsing, and have the option of reverting to the standard architecture if that seemed appropriate. In particular, this implied that source-language semantic representations needed to be such that they could reasonably be produced using phrase-spotting techniques.

In the grammar-based version, speech recognition uses a set of CFG-based language models (one per subdomain), compiled, using the REGULUS 2 toolkit, from a single linguistically motivated unification grammar (Rayner *et al.*, 2003b; Regulus, 2005). This makes it possible to support efficient structure-sharing between many similar subdomains with overlapping vocabulary and structure. Each subdomain-specific grammar is defined by a small training corpus, typically containing 500 to 1000 examples. The same corpus material is also used to perform probabilistic tuning of the resulting CFG language model. The statistical/robust version uses a normal class N-gram language model built using the Nuance SayAnything© package, together with a set of phrase-spotting rules. (Rayner *et al.*, 2004) reports experiments in which we compare performance for the two different versions of the system.

Both versions of the system use the same translation engine. Translation is interlingual and rule-based. Target language generation is also performed using suitably compiled linguistically motivated unification grammars. Output speech is produced using either a commercial TTS engine or concatenated recorded wavfiles, depending on the language.

3 Translating temporal and causal constructions

Initial versions of the MedSLT system (Rayner *et al.*, 2003a) used a completely flat representation format and a transfer-based translation architecture. For example, the English query "does the pain radiate to the jaw?" was represented as

```
[[utterance_type,ynq],[symptom,pain],[state,radiate],
 [tense,present],[prep,to_loc],[body_part,jaw]]
```

The Japanese translation "ago made itami wa hirogarimasu ka" (jaw-to pain-TOPIC radiate-POLITE-PRES Q) is represented as

```
[[utterance_type,sent],[symptom,itami],[state,hirogaru],
 [tense,present],[prep,made],[body_part,ago]]
```

When the scheme works, as it does here, the advantages are apparent: although the source and target versions have fairly different syntactic structures, the elements of the flat representations are in one-to-one correspondence. Transfer can be effected in a straightforward compositional fashion, and the constrained nature of the domain ensures that only valid target language translations can be produced from the target representation.

Problems arise, however, for causal and temporal constructions. For structurally similar languages, the same kind of solution tends to work reasonably well. For example, the representation of the English query "is the pain aggravated by coughing?" is

```
[[utterance_type,ynq],[symptom,pain],[event,aggravate],
 [tense,present],[cause,coughing]]
```

This can be translated into French as "la douleur est-elle aggravée par la toux?", which is represented similarly as

```
[[utterance_type,ynq],[symptom,douleur],[event,aggraver],
 [tense,present],[cause,toux]]
```

For unrelated language-pairs, this kind of solution is much more problematic. Although a literal translation of "is the pain aggravated by coughing?" into Japanese is not completely impossible, natural translations will not use a verbal construct corresponding to "aggravate", or a nominal construct corresponding to "coughing". It is instead preferable to use a subordinating conjunction construction, for example "seki wo suru to itami wa hidoku narimasu ka" (cough-OBJ make when pain-TOPIC worse become Q)[3].

Examples like these create a dilemma. Flat key/value representations are very suitable for robust phrase-spotting architectures, but there is no good way to handle a construction like a subordinate clause using a flat representation; both syntactically and semantically, a subordinate clause is clearly a nested structure. Unfortunately, the nature of the medical diagnosis domain means that temporal and causal constructions are extremely common. We have already seen "aggravate"; other typical examples are "relieve" ("does massage relieve the headache?"), "cause" ("is the headache caused by stress?"), "precede" ("is the headache preceded by nausea?") and "associated with" ("is the headache associated with vomiting?"). In English, too, natural phrasing often requires use of a subordinating conjunction. Although it is possible to say "is the pain relieved by lying down?", many people would prefer "is the pain better when you lie down?"

When we realised how important these phenomena were, our first reaction was to conclude that flat feature/value representations were simply inappropriate to a domain as complex as medical diagnosis questions: perhaps it was necessary to use general nested representations instead. If this were true, it would greatly complicate implementation of both the speech understanding and translation components of the system.

Further analysis, however, convinced us that this view of the situation was too extreme, and that a sensible compromise solution existed between the opposing positions of flat feature/value lists and general nested structures. Most importantly, we can in the context of this task reduce all temporal and causal relationships to one of the following canonical schemas: (1) [WHEN] Clause1 WHEN Clause2; (2) [BEFORE] Clause1 BEFORE Clause2; (3) [AFTER] Clause1 AFTER Clause2.

Figure 1 shows examples of how different concepts can be paraphrased in this way. Our new strategy then became the following: move to an interlingual translation architecture, and use the canonical versions of the temporal and causal relations as the interlingual representation.

Note that we are in no way claiming that temporal and causal relationships can be conflated in general; in many other contexts, we would certainly have to distinguish them. What we

[3]In practice, "itami wa" (pain-TOPIC) would often be omitted, since the topic is clear from context.

is the headache **aggravated** by bright light? →
headache **is worse WHEN** you are exposed to bright light?

does massage **relieve** the headache? →
headache **is better WHEN** you receive massage?

does stress **give** you headaches? →
you **have** headache **WHEN** you are stressed?

is the headache **associated with** vomiting? →
you vomit **WHEN** you **have** a headache?

is the headache **accompanied** by nausea? →
you experience nausea **WHEN** you **have** a headache?

is the headache **preceded** by scintillations? →
you experience scintillations **BEFORE** you **have** a headache?

do you get headaches **after** a large meal? →
you have headache **AFTER** you eat a large meal?

Figure 1: Examples of reducing causal and temporal concepts to canonical form

are doing, rather, is exploiting the constraints of the medical diagnosis task to simplify the semantic representation language. In this very specific context, the justification for replacing causal questions with temporal ones is that the patient will not normally know what causes the symptoms, even if they believe they do — they only know about the temporal sequence of events. For this reason, the doctor will not receive misleading information if the patient answers the temporal question, irrespective of whether it was originally phrased as temporal or causal.

At the level of concrete representations, we conservatively extend the representation language by allowing one level of nesting in the key/value lists, so as to make it possible to represent the subordinate clause construction. Thus for example the representation of "do you have headaches when you drink coffee?" is

```
[[utterance_type,ynq],[pronoun,you],[state,have_symptom],
 [tense,present],[symptom,headache],[sc,when],
 [[clause,[[utterance_type,dcl],[pronoun,you],
  [action,drink],[tense,present],[cause,coffee]]]]
```

We have carefully chosen the above example so that the source and interlingua representations are in this case identical; in other words, "do you have headaches when you drink coffee?" is the canonical way to say this question. Following Figure 1, we design the rules which map source language representations to interlingua so that we get the same interlingual form for other phrasings of the same question. For example, "are your headaches caused by coffee?", with source representation

```
[[utterance_type,ynq],[symptom,headache],[substance,coffee],
 [event,cause],[tense,present]]
```

and "does coffee give you headaches?", with source representation

```
[[utterance_type,ynq],[symptom,headache],[substance,coffee],
 [event,give],[tense,present]]
```

will both yield the same interlingual form as "do you have headaches when you drink coffee?"

In order to realise the scheme we have just sketched out, we had to solve two main technical problems. First, we needed to be able to produce nested source language representations for utterances containing subordinate clauses. Second, we required a clean way to structure the rules which map source language representations into interlingual ones. We consider these two sets of issues separately.

3.1 Producing nested source language representations

Producing nested representations in the grammar-based version of the recogniser is straightforward: these can be built up in the usual way using compositional semantics. The challenge is to produce them in the version of the system which uses statistical recognition, where we are limited to robust surface processing on a noisy recognition string. This section briefly describes our implemented solution.

Processing consists of three phases. First, a set of rules is applied that attempts to detect start- and end-boundaries for subordinate clauses. A typical rule in this group[4] is

```
boundary([when],[not_word(do/does/have/has/can)],start).
```

This guesses the start of a subordinate clause after the word "when", and before a word that is not one of the words "do", "does", "have", "has" or "can".

Once the recognition string has been segmented into clauses, a second set of rules is applied, to guess key/value pairs. A typical rule in the second group is

```
pattern([lean/leaning,forward],[action,lean_forward]).
```

This guesses the key/value pair `[action,lean_forward]` if a sequence is found consisting of the word "lean" or "leaning" followed by the word "forward".

Finally, a set of post-processing rules is applied, which fills in default values for unset features in the representation of each clause. For example, `tense` is by default set to `present`, and `utterance_type` to `ynq` if a verb is present, and `phrase` otherwise.

3.2 Mapping temporal and causal concepts into canonical form

Translation rules in the MedSLT system are implemented using the Prolog-based formalism defined by the Regulus toolkit (Rayner *et al.*, 2005). Basically, this allows definition of rules mapping lists of key/value pairs to lists of key/value pairs. Lists can optionally contain up to

[4]The form of the rules has been simplified slightly for presentational purposes.

one level of nesting, using the `[clause, ...]` representation of subordinate clauses shown above. Rules may be conditional on the presence or absence of partially specified elements in the rest of the list; there is also support for use of macros. Macros may be non-deterministic, in which case the rule expands into multiple copies. All these features are illustrated in the following (artificial) example,

```
transfer_rule([[polarity,OnOff]], [[event, @onoff(OnOff)]])
   :- context([event,switch]).

macro(onoff(on), switch_on).
macro(onoff(off), switch_off).
```

This says that the lists `[[polarity,on]]` and `[[polarity,off]]` are respectively mapped to the lists `[[event,switch_on]]` and `[[event,switch_off]]` in a context which also contains the key/value pair `[event,switch]`.

We now describe how we realise within this framework the types of transformation informally sketched in Figure 1. Comparing the left- and right-hand sides of the examples, we can see that the changes involved are of two types. On the one hand, the portions marked in bold pick out transformations associated with causal relations. Thus, informally, we transform "aggravated by" into "is worse when". Alongside these, we have transformations which map nominal concepts into associated verbal counterparts; so, again informally, we map "bright light" into "be exposed to bright light". The problem we need to solve here is how to structure the rule-base so that these two groups of rules can be kept separate and thus orthogonal.

The solution we have implemented is to realise the nominal-to-verbal transformations as macros, and the causal-to-temporal transformations as rules using those macros. The following typical rule (slightly simplified) handles a transformation which could be informally described as "X causes (symptom)" to "you have (symptom) when X occurs":

```
transfer_rule(
    [Noun,[event,cause]],
    [[state,have_symptom],[sc,when],
     [clause,
      [[utterance_type,dcl],[pronoun,you],[tense,present],
        @causal_noun_to_vp(Noun)]]])
  :- context([symptom,_]).
```

This operates in a context where there is an element matching `[symptom,_]` in the environment. The left-hand side is a list consisting of `[event,cause]` together with a causal noun; the right-hand side is a list containing the key/value pairs `[state,have_symptom]`, `[sc,when]`, and a subordinate clause where the subject is "you" and the verb-phrase is the verbal counterpart of the noun on the left-hand side.

The non-deterministic macro `causal_noun_to_vp` contains one definition for each causal noun. Typical entries are

```
macro(causal_noun_to_vp([[substance,tea]]),
       [[action,drink],[substance,tea]]).
```

```
macro(causal_noun_to_vp([[substance,large_meal]]),
      [[action,eat],[substance,large_meal]]).
macro(causal_noun_to_vp([cause,massage]),
      [[state,experience],[cause,massage]]).
```

The first two entries are obvious: the nominal concepts "tea" and "large meal" map into the verbal concepts "drink tea" and "eat large meal". The third entry shows another common pattern. In many cases, the verb associated with the nominal concept is semantically neutral; we represent this using the key/value pair `[state,experience]`. The rules which map from the interlingua to the target language may give `[state,experience]` a more specific lexical realisation. Thus for example when moving from interlingua to English we map `[[state,experience], [cause,massage]]` to "receive massage"; if the target language is Japanese, it is mapped to the neutral "massaaji suru" (do massage).

Although this representational scheme is quite simple, we have been surprised to see what a wide range of complex translation mismatches it can handle. One particularly interesting case concerns WH-pronouns. These are represented similarly to other causal concepts, so for example "what relieves your headaches?" is represented as

```
[[utterance_type,whq],[spec,what],[event,relieve],
 [tense,present],[symptom,headache]]
```

We map this into interlingual form by simply adding another definition of `causal_noun_to_vp`,

```
macro(causal_noun_to_vp([spec,what]),
      [[state,experience],[spec,what]]).
```

For Japanese, `[[state,experience], [spec,what]]` can be mapped directly into the expression "nani wo suru" (do what); thus we translate "what relieves your headaches?" into the quite natural "nani wo suru to zutsu ga osamarimasu ka" (what-OBJ do when headache-SUBJ get-better-Q). A detailed evaluation of the performance of the system can be found in (Rayner *et al.*, 2004).

4 Summary and conclusions

We have presented an overview of MedSLT, a medium vocabulary medical speech translation system, focussing on the representational issues that arise when translating temporal and causal concepts. Although flat key/value structures are strongly preferred as semantic representations in speech understanding systems, we argue that it is infeasible to handle the necessary range of concepts using only flat structures.

By exploiting the specific nature of the task, we have shown that it is possible to implement a solution which only slightly extends the representational complexity of the semantic representation language, by permitting an optional single nested level representing a subordinate clause construct. We have sketched our solutions to the key problems of producing minimally nested representations using phrase-spotting methods, and writing cleanly structured rule-sets that map temporal and phrasal representations into a canonical interlingual form.

References

BLACK A., BROWN R., FREDERKING R., SINGH R., MOODY J. & STEINBRECHER E. (2002). TONGUES: Rapid development of a speech-to-speech translation system. In *Proceedings of HLT: Human Language Technology Conference*.

KNIGHT S., GORRELL G., RAYNER M., MILWARD D., KOELING R. & LEWIN I. (2001). Comparing grammar-based and robust approaches to speech understanding: a case study. In *Proceedings of Eurospeech 2001*, p. 1779–1782, Aalborg, Denmark.

LAVIE A., LANGLEY C., WAIBEL A., PIANESI F., LAZZARI G., COLETTI P., TADDEI L. & BALDUCCI F. (2001). Architecture and design considerations in NESPOLE!: a speech translation system for e-commerce applications. In *Proceedings of HLT: Human Language Technology Conference*, San Diego, California.

LAVIE A., WAIBEL A., LEVIN L., FINKE M., GATES D., GAVALDA M., ZEPPENFELD T. & ZHAN P. (1997). JANUS-III: Speech-to-speech translation in multiple languages. In *Proceedings of the 6th European Conference on Speech Communication and Technology (EuroSpeech99)*, p. 99–106.

NUANCE (2003). http://www.nuance.com. As of 25 February 2003.

PHRASELATOR (2004). http://www.phraselator.com. As of 8 Dec 2004.

RAYNER M. & BOUILLON P. (2002). A phrasebook style medical speech translator. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (demo track)*, Philadelphia, PA.

RAYNER M., BOUILLON P., HOCKEY B., CHATZICHRISAFIS N. & STARLANDER M. (2004). Comparing rule-based and statistical approaches to speech understanding in a limited domain speech translation system. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation; also ftp://issco-ftp.unige.ch/pub/publications/tmi_045.pdf*, Baltimore, MD.

RAYNER M., BOUILLON P., VAN DALSEM V., HOCKEY B., ISAHARA H. & KANZAKI K. (2003a). A limited-domain English to Japanese medical speech translator built using REGULUS 2. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (demo track)*, Sapporo, Japan.

M. RAYNER, D. CARTER, P. BOUILLON, V. DIGALAKIS & M. WIRÉN, Eds. (2000). *The Spoken Language Translator*. Cambridge University Press.

RAYNER M., HOCKEY B. & BOUILLON P. (2005). *Using Regulus*. http://cvs.sourceforge.net/viewcvs.py/regulus/Regulus/doc/RegulusDoc.htm. As of 30 January 2005.

RAYNER M., HOCKEY B. & DOWDING J. (2003b). An open source environment for compiling typed unification grammars into speech recognisers. In *Proceedings of the 10th EACL (demo track)*, Budapest, Hungary.

REGULUS (2005). http://sourceforge.net/projects/regulus/. As of 30 January 2005.

W. WAHLSTER, Ed. (2000). *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer.

YOUNG S. (2002). Talking to machines (statistically speaking). In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*, p. 9–16, Denver, CO.