Challenges in Adapting an Interlingua for Bidirectional English-Italian Translation

Violetta Cavalli-Sforza[1], Krzysztof Czuba[2], Teruko Mitamura[2], and Eric Nyberg[2]

[1] San Francisco State University, Department of Computer Science
1600 Holloway Avenue, San Francisco, CA 94132
vcs@Sfsu.edu
[2] Language Technologies Institute, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213
{kczuba, teruko, ehn}@cs.cmu.edu

Abstract. We describe our experience in adapting an existing high-quality, interlingual, unidirectional machine translation system to a new domain and bidirectional translation for a new language pair (English and Italian). We focus on the interlingua design changes which were necessary to achieve high quality output in view of the language mismatches between English and Italian. The representation we propose contains features that are interpreted differently, depending on the translation direction. This decision simplified the process of creating the interlingua for individual sentences, and allows the system to defer mapping of language-specific features (such as tense and aspect), which are realized when the target syntactic feature structure is created. We also describe a set of problems we encountered in translating modal verbs, and discuss the representation of modality in our interlingua.

1 Introduction

In this paper, we describe our experience in adapting an existing high-quality, interlingual, unidirectional machine translation system for a new domain and bidirectional translation. We concentrate on some of the changes in the interlingua design that were necessary to ensure high quality output.

KANT [7] is an interlingua-based software architecture for knowledge-based machine translation. The CATALYST project used the KANT technology for translation of technical documentation in the domain of heavy equipment from English to several European languages. At present, systems for translation to Spanish, French and German are in production use, and a Portuguese system for the same domain is almost fully developed. Prototypes of varying coverage for different domains have been developed for languages as diverse as Italian, Chinese, Japanese, Turkish, and Arabic [2], [6], [4].

In the CATALYST system, translation is unidirectional, from English to other languages, and not all of the target languages were known before the interlingua was designed. Although the interlingua design does represent the meaning of the input by abstracting away from the surface details, it is somewhat isomorphic to the semantic structure of English [3], [1].

In this paper we discuss our experiences in the MedTran project, an application of KANT technology to bidirectional English-Italian translation of medical records. The goal of the project was to facilitate communication between monolingual physicians and medical staff working in an international facility. Very little source material was available in Italian. For English we had access to a sizeable corpus of transcribed documents. The input was not controlled, could be ungrammatical, and might contain structures which are not part of common written English. Ambiguous constructions might require interactive disambiguation to promote high-quality translation. In addition, the medical domain emphasizes different concepts and constructions than the domains for which the KANT interlingua was designed originally.

We began by implementing a proof-of-concept demonstration system that translated from English into Italian. Building the prototype forced us to focus on a number of linguistic issues that were not as significant in the original KANT domains. Based on this experience, we redesigned the interlingua, taking into consideration specific issues that arose when translating from English into Italian and vice-versa. We review briefly a few of the key findings, sketch our approach, and draw some general conclusions from our experience.

2 System Architecture

The general architecture of KANT is shown in Fig. 1 . Translation is performed on one sentence at a time, with separate analysis and generation phases.

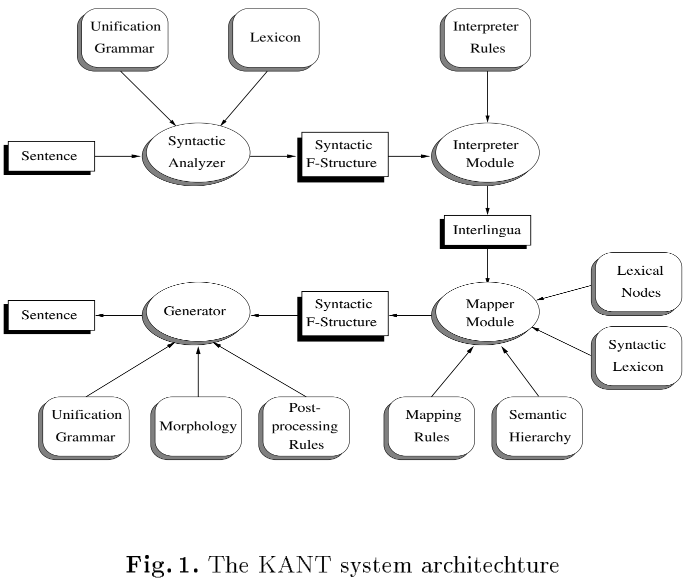

Fig. 1. The KANT system architechture

During the analysis phase, each sentence is first converted into tokens. Using a lexicon enriched with syntactic information, a morphological analyzer, source language grammar rules, and optionally semantic information, the tokenized sentence is parsed into a feature structure (FS), a recursive list of feature-value pairs that reflects the syntactic structure of the input. Using a set of analysis mapping rules, the interpreter converts the FS into a tree-structured interlingua representation (IR), which abstracts away many of the syntactic details of both source and target language, while conveying the meaning of the source [3].

In the generation phase, generation mapping rules convert the IR into a FS that reflects the syntactic structure of the target language. The mapper recursively traverses the IR, converting subtrees to FS constituents and terminal concept nodes to target lexical items. The mapper uses general mapping rules, bilingual data structures called 'lexical nodes', and a syntactic lexicon for the target language. The FS is then processed by the generator, which uses target language grammar rules and morphological generation rules to produce a preliminary target language output. In the final generation step, post-processing rules clean up spacing and punctuation, and handle surface-level issues such as elision and contraction of words.

3 Input Characteristics

The input in the MedTran project consists of various kinds of notes made by physicians (progress notes, discharge summaries, radiology reports, etc.), traditionally dictated and transcribed. For the development of the proof-of-concept prototype system we used five examples of medical texts drawn from two document types, discharge summaries and progress notes, for a total of approximately 110 distinct phrases and sentences. The texts are semi-structured, with labels usually identifying different sections of the text (e.g., *Cardiovascular*, *Respiratory*, *Renal*). This structure reflects the sequence in which medical examinations are performed and can be used for disambiguation purposes in the MT system.

On the linguistic side, the texts included idioms that could not be translated literally and required some restructuring, as illustrated in Example 1.

Example 1.
(a) This is a <u>60 year old male</u> who <u>is</u> end stage liver disease.
(b) *Questo è un <u>maschio di 60 anni</u> che <u>è affetto da</u> malattia epatica terminale.*
(c) This is a <u>male of 60 years</u> who <u>is affected by</u> end stage liver disease.

In this example, the English phrase "end stage liver disease" (*malattia epatica terminale*) is used as a predicate in (a). In the Italian translation, shown in (b), the sentence must be changed to a passive construction. The literal word-by-word translation into English of the Italian output is shown in (c). The underlined segments of the sentences show the effect of the required restructuring.

Since the texts were mostly dictated by non-native speakers of English, minor adjustments to the input were required in a small number of cases in which the input was not only ungrammatical but also difficult to understand even for a human reader. Other changes we made were in the use of tenses (e.g., *The liver enzymes continued to be rising.*), prepositions, and punctuation.

Another characteristic of the texts was ample use of a range of tenses, temporal modifiers and other expressions implying the time dimension. Time turned out to be an important issue in interlingua design for the medical domain, whereas it had been less important in previous applications of the KANT technology.

4 Interlingua Design: Issues and Approach

The KANT interlingua representation (IR) contains features that are purely semantic and features that are primarily grammatical in nature. In some cases, grammatical features are used by the generation module to produce a maximally accurate target translation. In the MedTran system, the IR also contains grammatical features. However, because translation is bidirectional, and English and Italian differ in significant ways, the information that must be represented in the IR might not be identical for the two languages. This issue arose in the representations of mood, tense, aspect, and modality. We addressed this by choosing a set of IR features that are shared by the two languages, but that are interpreted somewhat differently during translation. In some cases, the same features are interpreted differently because they are used to represent different linguistic concepts. In other cases, one language uses only a subset of the features, or of the values defined for a particular feature. Examples of all cases are given below. With this approach, the IR encodes as much information as possible from the source language, allowing the generation phase to use this information as needed to produce an appropriate target language output. Another important modification is the introduction of new features to better capture the use of temporal and location modifiers, which must be generated with special care in domains where time is an important component of the information conveyed by the text. This section describes the challenges we encountered in the abovementioned areas, and sketches the solutions we developed.

4.1 Verb Mood, Tense and Aspect

Although the tense and aspect systems of English and Italian show some similarities, there are also many differences that make mapping tenses between the languages difficult in an MT system.

The feature verb-mood (with values subjunctive, conditional, indicative, imperative, infinitive, gerund and participle), is used to represent the mood of the verb. The feature verb-mood is distinct from the feature sentence-mood (with values declarative, interrogative, exclamative, and imperative). In Italian, it is possible for an imperative sentence to use different verb moods, for example: a subjunctive, to indicate a formal command; an imperative, for informal commands; or an infinitive, common in product instructions and manuals.

The tense and aspect system also differs between the two languages. In English, most verbs can be marked independently for the progressive and perfective aspect and for tense. For example, "he examines" is a simple present, neither perfective nor progressive; "he has been examining" is both perfective and progressive; "he is examining" is only progressive; and "he has examined" is only perfective. The same combinations of perfective and progressive exist for the past and the future tenses. To encode this information, when translating from English to Italian, the IR uses the features tense, with values present, past, and future, and the features perfective, and progressive with the values + and -.

In Italian, the distribution of aspect is not entirely independent of the distribution of tense, especially with respect to the expression of progressive aspect. The indicative mood has eight tenses, four simple tenses (present, past, imperfective, future), and four compound tenses (explained below). The subjunctive mood has two simple tenses (present and imperfective) and two compound tenses. The conditional mood has only one simple and one compound tense. In each mood, compound tenses are formed by using an auxiliary verb (normally *avere* "to have" or *essere* "to be") in one of the mood's simple tenses followed by a past participle. The IR features tense and perfective are used differently when going from Italian to English. A simple verb uses only the tense feature, with the same values as English plus the value imperfective. A compound verb is encoded using tense to capture the tense of the auxiliary verb, plus the feature perfective set to +. For example, *avesse mangiato* (roughly "that he had eaten"), an example of the Italian subjunctive compound tense pluperfect, uses the imperfective of "to have" (*avesse*) and the past participle of *mangiare* ("to eat"); it would be encoded as (verb-mood subjunctive), (tense imperfect), (perfective +).

The relationship of progressive aspect in English and Italian is complex. The indicative imperfective tense is sometimes used to convey habitual or repetitive actions in the past – where English might use a past progressive form or a different construction (Example 2) – or an ongoing action – where English might use a progressive or a simple past (Example 3). The verb ending *-ava* is a third person singular masculine or feminine indicative imperfective ending.

Example 2.
Fumava un pacchetto di sigarette al giorno.
S/he was smoking/used to smoke a pack of cigarettes a day.

Example 3.
Parlava mentre il dottore lo visitava.
He was talking/talked while the doctor was visiting/visited him.

The imperfective is also used to express enduring states in the past (Example 4), whereas a simple or compound past is used with point events (Example 5).

Example 4.
Ieri la temperatura era 37,5.
Yesterday the temperature was 37.5.

Example 5.
Ieri le entrate sono state 3,4 litri e le uscite sono state 4,6 litri.
Yesterday input was 3.4 liters and output was 4.6 liters.

Progressive constructions in Italian are used to convey being in the process of acting or experiencing and are used more rarely than in English. They are formed using the verb *stare* (literally 'to stay') as an auxiliary and the gerund of the main verb, as in Example 6: *stiamo* is the present indicative first person plural of *stare*, *dando* is the gerund of *dare* ("to give").

Example 6.
Gli stiamo dando insulina.
We are giving him insulin.

The progressive feature with value + is used exclusively to indicate this type of construction. The tense feature indicates the tense of the auxiliary *stare*, which can only be a simple tense. Hence, in representing an Italian input, the features progressive and perfective never co-occur.

4.2 Modals

The design of the IR for modal verbs was one of the more complex issues we had to address. English modals (e.g., "may", "can", "should") are not full-fledged verbs. In Italian, however, modal verbs (e.g., *potere*, *dovere*) are fully conjugated verbs that require a non-perfective or perfective infinitive form of the main verb. Because of these differences, and the ambiguity generated by the use of modals, the IR captures the value(s) of modality expressed by the modal verb in a special feature. It also captures tense, mood, and aspect information present in the input for both main and modal verbs. Translation from English uses a subset of the features used required by Italian.

The mandatory IR feature for representing modals is modality, which can take on several values. The modality habit encodes the modal "would" as used in the sentence "The patient would walk a few minutes every morning", which would be translated as an imperfective of the verb "to walk" (*passeggiava*). The modality hypothetical is used for the modal "should" in the sentence "Should the patient improve, we will decrease the dose", which would require a hypothetical construction with *se* ("if") in translation. The values ability, permission, and possibility are used for the modals "can" and "could", among others. The values expectation, necessity, obligation are used for the modal "should", among others. Since it is frequently difficult to distinguish among these modalities, and not always necessary for correct translation, they frequently occur as a disjunctive value (e.g., (:or necessity obligation)).

Other features may be present as well. The feature occurrence (with values certain, uncertain, or unrealized), combines with modality to encode different shades of modality. The features modal-tense (mostly for Italian) and modal-perfective (only for Italian), with values + and -, encode tense and aspect for the modal verb. The verb features described in Section 4.1 encode the main verb.

In the remainder of this section we provide a few examples of use of these features in both directions of translation.

The Modals "can" and "could". The modals "can" and "could" are responsible for much ambiguity in English. The uses of "can" in Example 7 cannot be disambiguated without extensive semantic analysis; however they can all be translated into Italian with the verb *potere* followed by a non-perfective infinitive, maintaining the same ambiguity. In these cases the IR would have (modality (:or ability permission possibility)).

Example 7.
The patient can open his eyes. (ability)
The patient can go home tomorrow. (permission)
The tumor can metastasize. (possibility)

Uses of "could", however, must be disambiguated in order to produce a correct translation.

"Could", used as the past of "can", as in the example "We could not identify the source of the bleeding", is translated with an appropriate past tense of *potere*. The IR uses (modality (:or ability permission possibility)) and (modal-tense past).

"Could", followed by a non-perfective infinitive, as in Example 8, may express higher uncertainty than "can" and is encoded in the IR with (occurrence uncertain). The Italian translation requires the present conditional of *potere* followed by a non-perfective infinitive.

Example 8.
Questo potrebbe essere dovuto a un'epatite.
This could be due to hepatitis.

"Could", followed by a perfective infinitive (encoded as (perfective +)), in some contexts expresses uncertainty in the past, encoded with (occurrence uncertain). It must be translated into Italian with the present conditional of *potere* (*potrebbe*) followed by a perfective infinitive (e.g. *avere avuto*, from *avere* "to have"), as in Example 9.

Example 9.
Potrebbe avere avuto un piccolo infarto.
He could have had a small infarction.

In other contexts, "could" followed by a perfective infinitive expresses unrealized ability, possibility or permission, which is encoded as (occurrence unrealized). This is translated in Italian with a past conditional of *potere* (*avremmo potuto*) followed by a non-perfective infinitive (e.g. *operare* "to operate"), as in Example 10.

Example 10.
Avremmo potuto operare ieri.
We could have operated yesterday (but we didn't).

The Modal *Dovere* in Italian. Ambiguity in modal use is possible in translating from Italian to English as well. The modal *dovere* in Italian conveys expectation, necessity and obligation. While obligation and necessity modalities do not need to be distinguished from each other, in some cases they need to be distinguished from expectation, as shown in Example 11 below. (*Dobbiamo* and *devono* are present indicative first and third person plural respectively of *dovere*).

Example 11.
Dobbiamo informare la famiglia. (obligation)
We must inform the family.

Dobbiamo operare il paziente in settimana. (necessity)
We must operate the patient within the week.

I risultati devono arrivare in settimana. (expectation)
The results will/should arrive this week.

As an example of modal encoding in the IR when translating from Italian, the past conditional modal *avremmo dovuto* in Example 12 is represented by (modal-tense present), (modal-perfective +), (occurrence unrealized).

Example 12.
Avremmo dovuto operare prima.
We should have operated earlier (but we didn't).

4.3 Time and Location Modifiers

Modifier Positioning. The IR design considers a number of clausal and sentential modifiers, including subordinate clauses, adjoined modifiers (e.g., *if necessary*), discourse markers (e.g., *actually, nonetheless*), adverbial phrases, noun phrases, and prepositional phrases. The IR for modifiers includes the feature position, which can take on at least the values initial, if the modifier occurs at the beginning of the clause, and end if it occurs at the end of the clause. For some kinds of modifiers, other internal positions are also possible in English, depending on the presence of auxiliary verbs and other syntactic characteristics, as illustrated by Example 13 [8]. A similar range of options is available for modifier positioning in Italian.

Example 13.
By then, the patient should have been feeling better. (initial)
The patient, by then, should have been feeling better. (initial-medial)
The patient should, by then, have been feeling better. (medial)
The patient should have, by then, been feeling better. (medial-medial)
The patient should have been, by then, feeling better. (end-medial)
The patient should have been feeling, by then, better. (initial-end)
The patient should have been feeling better, by then. (end)

While it is not strictly necessary to generate modifiers in all possible positions, in order to obtain more faithful translations it is important to record modifier position on input and choose the position on output accordingly. In Italian, as in English, initial positioning is more "neutral", while modifiers positioned at the end of a sentence carry more emphasis and modifiers positioned right after the subject are somewhat parenthetical. Time expressions, in particular, appear to be widely used in the medical domain. Their positioning can carry subtle shades of meaning. Consider the following sentences:

Example 14.
(a) Tomorrow Dr. Boyle will request chest X-rays for Mr. Smith.
(b) Dr. Boyle will request chest X-rays for Mr. Smith tomorrow.
(c) Dr. Boyle, tomorrow, will request chest X-rays for Mr. Smith.

In Example 14, the initial position of "tomorrow" in (a) is unremarkable, while the end position in (b) emphasizes the adverbial and suggests a contrast with another time, for example "later today". In (c), "tomorrow" is added almost as an afterthought.

In our sample input, mostly dictated by non-native speakers of English, temporal modifiers were sometimes placed a little anomalously. To remain as faithful as possible to the original input, we did not correct positioning in generation unless it violated positioning rules for the target language. For example, in English an adverbial modifier cannot be positioned between a verb and its direct object, but it can follow the direct object. In Italian both positions are often possible, but positioning between the verb and the direct object is preferable.

Example 15.
L'infermiera chiamò immediatamente il Dott. Boyle.
*The nurse called immediately Dr. Boyle.

L'infermiera chiamò il Dott. Boyle immediatamente.
The nurse called Dr. Boyle immediately.

Co-positioning of Temporal and Location Modifiers. We found it desirable to keep time and location modifiers together but separate from other types of modifiers. Isolating them gives greater control over their positioning. Joint handling is motivated by the observation that they often appear together in the input and are frequently related. Extracting time modifiers and positioning them separately from location modifiers can easily lead to breaking subtle but important ordering relationships. Consider the following examples.

Example 16.
(a) Tomorrow at the hospital, Dr. Boyle will visit the patient.
(b) Dr. Boyle will visit the patient tomorrow at the hospital.
(c) Dr. Boyle will visit the patient at the hospital tomorrow.
(d) Tomorrow, Dr. Boyle will visit the patient at the hospital.
(e) At the hospital, Dr. Boyle will visit the patient tomorrow.

In (a) there is more emphasis on the visiting, in (b) on the time and place. In (b) and (d) there is more emphasis on the place, in (c) on the time, while (e) is somewhat anomalous. The combination of placing the time and location modifiers in the same IR slot, recording their position, and keeping them in the same relative order in which they appeared in the source sentence, facilitates generating them in the correct place and order in translation.

5 Conclusions

In this paper we have described some of our experiences with adapting an interlingua representation for a unidirectional MT sytem when moving to bidirectional translation between English and Italian and a different domain. While, by definition, an interlingua representation abstracts away the syntactic details of the source language, an effective interlingua may need to represent grammatical information present in the input if this information captures important semantic and functional distinctions that are made by each language. We took the approach of using a common set of features for the two languages, which allows the representation to be language independent, as an interlingua should be. At the same time, we allowed the features to be used differently or not used at all depending on the direction of translation. This approach allows us to capture specific feature sets which more accurately represent degrees of meaning in the source language. The finer detail can be utilized by specific target language generators to produce more accurate translations when the target language supports the same feature set.

References

1. Czuba, K., Mitamura, T., and Nyberg, E.: Can practical interlinguas be used for difficult analysis problems? In: Proceedings of AMTA-98 Interlingua Workshop (1998)
2. Hakkani, D., Tür, G., Oflazer, K., Mitamura, T., and Nyberg, E.: An English-to-Turkish interlingual MT system. In: Proceedings of AMTA-98 (1998)
3. Leavitt, J., Lonsdale, D., and Franz, A.: A reasoned interlingua for knowledge-based machine translation. In: Proceedings of CSCSI-94 (1994)
4. Li, T., Nyberg, E., and Carbonell, J.: Chinese sentence generation in a knowledge-based machine translation system. Technical Report CMU-CMT-96-148, Carnegie Mellon University (1996)
5. Mitamura, T., and Nyberg, E.: Controlled English for knowledge-based MT: Experience with the KANT system. In: Proceedings of TMI-95 (1995)
6. Nyberg, E., and Mitamura, T.: A real-time MT system for translating broadcast captions. In: Proceedings of MT Summit VI (1997)
7. Nyberg, E., and Mitamura, T.: The KANT system: Fast, accurate, high-quality translation in practical domains. In: Proceedings of COLING-92 (1992)
8. Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J.: A Comprehensive Grammar of the English Language Longman, London New York (1985)