An English-to-Turkish Interlingual MT System

Dilek Zeynep Hakkani[1], Gökhan Tür[1], Kemal Oflazer[1], Teruko Mitamura[2], and Eric H. Nyberg, 3rd[2]

[1] Department of Computer Engineering and Information Science, Bilkent University, Ankara 06533, Turkey

[2] Center for Machine Translation, Carnegie Mellon University, Pittsburgh PA 15213, USA

Abstract. This paper describes the integration of a Turkish generation system with the KANT knowledge-based machine translation system to produce a prototype English–Turkish interlingua-based machine translation system. These two independently constructed systems were successfully integrated within a period of two months, through development of a module which maps KANT interlingua expressions to Turkish syntactic structures. The combined system is able to translate completely and correctly 44 of 52 benchmark sentences in the domain of broadcast news captions. This study is the first known application of knowledge-based machine translation from English to Turkish, and our initial results show promise for future development.

1 Introduction

This paper describes the integration of a Turkish generation system [2], developed in the framework of an ongoing large-scale research project on Turkish natural language processing, with the KANT knowledge-based machine translation system, developed under the KANT project at Carnegie Mellon University's Center for Machine Translation [7]. The result is a prototype English–Turkish, interlingua-based machine translation system. In order to integrate these independently developed systems, we have designed and implemented a mapping module, using the KANT mapper software developed [4], which transforms the interlingua representation of each sentence to a feature structure (hereafter, f-structure) for Turkish; the resulting Turkish f-structure is then input to the existing Turkish sentence generator, producing Turkish surface forms.

The paper is structured as follows: In Section 2 we give a brief introduction to relevant features of Turkish. In Section 3, we briefly present the architecture and details of the Turkish subsystem comprising the mapper and the Turkish generator. We then present some experimental results, and discuss a set of important issues that we encountered during the design and implementation of the system.

2 Turkish

Morphologically, Turkish is an agglutinative language, with very productive inflectional and derivational suffixation processes by which it is possible to generate

thousands of forms from a given root word. A slightly exaggerated example of a Turkish word formation is illustrated by the following nominal:

(1) Ankaralılaştıramayabileceklerimiz
Ankara-lı-laş-tır-ama-yabil-ecek-ler-imiz
those whom we can not convert to a citizen of Ankara

Turkish morphotactics are finite-state, and the surface realization of words is constrained by morphographemic processes such as vowel harmony. For details regarding Turkish grammar and word formation rules, one may refer to Lewis [5]; see also Oflazer [10] for a finite-state description of Turkish morphology.

With respect to word order, Turkish can be considered a *subject-object-verb* (SOV) language, in which constituents can change order rather freely in almost all sentential constructions, depending on the constraints of text flow or discourse. The grammatical roles of constituents are identified by explicit morphological case markings rather than their constituent order. For example, the word 'masa' (table), case marked accusative, is a definite direct object. The same word, when case marked dative, expresses a goal (unless it is accompanied by an idiosyncratic verb which subcategorizes for a dative complement)[3,4]:

(2)a. Masa-yı sil-di-m
table-ACC wipe-PAST-1SG
'I wiped the table.'
b. Kitab-ı masa-ya koy-du-m
book-ACC table-DAT put-PAST-1SG
'I put the book on the table.'

Word order variation in Turkish is, for the most part, dictated by information structure constraints which capture and encode, to a certain extent, discourse-related factors [15].

3 The Architecture of the System

The system which generates Turkish sentences from interlingua representations consists of 4 subsystems: the mapping system, the sentence generation system, the interface, and the morphological generation system (see Figure 1).

To demonstrate the function of each component, we will use the example sentence:

"Tosco will become the nation's largest independent refinery."

[3] From this point on we will give Turkish forms with -'s indicating morpheme boundaries, where necessary.

[4] In the glosses, `3SG` and `1SG` denote third person singular and first person singular verbal agreement, `P3SG` denotes third person singular possessive agreement, `LOC, ABL, DAT, GEN, ACC` denote locative, ablative, dative, genitive, and accusative case markers, `PAST` denotes past tense, and `INF` denotes a marker that derives an infinitive form from a verb.

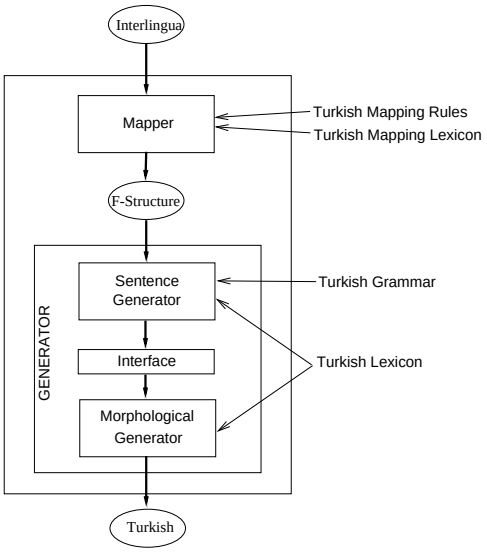

Fig. 1. The Turkish generation system.

The interlingua generated by the KANT analyzer is as follows:[5]

```
(*A-BECOME
 (PUNCTUATION PERIOD)
 (FORM FINITE)
 (TENSE FUTURE)
 (MOOD DECLARATIVE)
 (ARGUMENT-CLASS BENEFICIARY+GOAL)
 (BENEFICIARY (*PROP-TOSCO
               (NUMBER MASS)
               (IMPLIED-REFERENCE +)
               (PERSON THIRD)))
 (GOAL (*O-REFINERY
        (NUMBER SINGULAR)
        (REFERENCE NO-REFERENCE)
        (PERSON THIRD)
        (UNIT -)
        (POSSESSOR (*O-NATION
                    (NUMBER SINGULAR)
```

[5] Most of the linguistic features used in the KANT interlingua (e.g., punctuation, form, tense, mood, argument class, number, person) should be self-evident. Some other features are artifacts of KANT's evolution as a technical text system. The `IMPLIED-REFERENCE` feature is used for nouns (such as the proper noun in the example) which have implicit reference, although they are not marked with a determiner. To promote representational consistency, the same structure (`*G-COORDINATION`) is used whether or not an explicit conjunction (such as "and") appears; hence in the example the (`CONJUNCTION`) is `NULL`.

```
                (REFERENCE DEFINITE)
                (UNIT -)
                (PERSON THIRD)))
        (ATTRIBUTE
         (*G-COORDINATION
          (CONJUNCTION NULL)
           (CONJUNCTS
            (:MULTIPLE
             (*P-LARGE
              (COMPARISON MOST)
              (DEGREE SUPERLATIVE))
             (*P-INDEPENDENT
              (DEGREE POSITIVE)))))))))
```

3.1 The Mapping System

The mapping system produces f-structures for Turkish from the interlingua representations, using a set of mapping rules and a mapping lexicon. For the example interlingua given above, the mapping system produces the following Turkish f-structure:

```
((REL IS-A)
 (ARGUMENTS
  ((SUBJECT
    ((SPECIFIER
      ((DETERMINER ((DEFINITE +)))))
     (REFERENT
      ((AGR ((NUMBER SINGULAR)
             (PERSON 3)))
       (RPROPER +)
       (ARG ((CONCEPT "tosco")))))))
   (PRED-OBJ
    ((SPECIFIER
      ((DETERMINER ((DEFINITE +)))))
     (MODIFIER
      ((QUALITATIVE
        ((ELEMENT
          (*MULTIPLE*
           ((P-NAME "bUyUk")
            (INTENSIFIER ((DEGREE EN))))
           ((P-NAME "baGImsIz"))))
         (CONJ " "))))))
     (REFERENT
      ((SEM ((COUNTABLE +)))
       (ARG ((CONCEPT "rafineri")))
       (AGR ((NUMBER SINGULAR)
             (PERSON 3)))))
     (POSSESSOR
      ((REFERENT
        ((SEM ((COUNTABLE +)))
```

```
        (ARG ((CONCEPT "Ulke")))
        (AGR ((NUMBER SINGULAR)
              (PERSON 3)))))
      (SPECIFIER
       ((DETERMINER ((DEFINITE +)))))))))))
(VERB ((ROOT TO-BE)
       (SENSE POSITIVE)
       (TENSE FUTURE)))
(SPEECH-ACT DECLARATIVE)
(VOICE ((ACTIVE +)))
(PUNCTUATION PERIOD)
(S-FORM FINITE)
(CLAUSE-TYPE ATTRIBUTIVE))
```

A fully detailed discussion of the mapping system is beyond the scope of this paper; in the remainder of this section, we describe how basic sentential components are mapped.

Verb Form Mappings. In order to realize the surface form of a verb in Turkish, it is necessary to determine certain morphological features of the verb in addition to its root, such as voice, polarity, tense, aspect, mood, and agreement features. The voice and polarity information can be directly obtained from the interlingua. The agreement depends on the subject of the sentence. Aspect, mood and tense features depend on the tense, perfective, progressive, and conditional information in the interlingua.

Argument Mappings. In the interlingua representation there is a semantic argument class feature, which states the possible arguments each verb may take. While mapping the arguments of a sentence, this information is used to determine the counterpart of each argument in the f-structure for Turkish. But sometimes, this information may not be enough, in which case, the verb's subcategorization information, and the type of the sentence (e.g., predicative, existential, etc.) and the voice of the main verb in Turkish are also required.

Noun Phrase Mappings. Most of the features for a noun phrase in the interlingua, like definiteness, agreement, quantifier, quantity, and possessor, are directly mapped to their counterparts in the Turkish f-structure.

Prepositional Phrase Mappings. The prepositional phrases which are attached to a noun phrase in the interlingua are mapped to one of the modifiers or specifiers depending on the preposition. The ones which are attached to the verb are mapped to either an argument of the sentence, or a postpositional phrase in Turkish. This selection depends on the preposition in question, and also on certain semantic conditions.

3.2 Sentence Generation System

The sentence generation system was originally designed and implemented by Hakkani and Oflazer [2] for use in a prototype transfer-based human-assisted machine translation system from English to Turkish [14]. This component is implemented using the CMU-CMT Genkit system [13], and is based on a recursively-structured finite state machine (much like a recursive transition network) which handles the constituent order variations of Turkish, implemented as a right-linear grammar backbone.

The sentence generation system receives as input an f-structure representing the content of the sentence, where all lexical selections have been made, and produces as output an f-structure for each word of the sentence, encoding relevant abstract morphological features such as: *agreement, possessive,* and *case* markers for nominals and *voice, polarity, tense, aspect, mood,* and *agreement* markers for verbal forms, as well as markers for *all productive derivations.*

3.3 Morphological Generation System

Morphological realization has been designed and implemented by Oflazer [10] using an external morphological analysis/generation component. This component performs (i) concrete morpheme selection, dictated by the morphotactic constraints and morphophonological context, (ii) handles morphographemic phenomena such as vowel harmony, and vowel and consonant ellipsis, and (iii) produces an agglutinative surface form.

3.4 Interface

The main task of the interface is to collect and send the output of the sentence generation system (which are morphological feature structures), to the morphological generation system in the required format, and then print out the surface form of the translated sentence. For the example sentence at the beginning of this section, the interface produces the following from the output of the sentence generation system:

```
[[CAT=NOUN][ROOT=Tosco][TYPE=RPROPER][AGR=3SG][POSS=NONE][CASE=NOM]]
[[CAT=NOUN][ROOT=ülke][AGR=3SG][POSS=NONE][CASE=GEN]]
[[CAT=ADVERB][ROOT=en]]
[[CAT=ADJ][ROOT=büyük]]
[[CAT=ADJ][ROOT=bağımsız]]
[[CAT=NOUN][ROOT=rafineri][AGR=3SG][POSS=3SG][CASE=NOM]]
[[CAT=VERB][ROOT=ol][SENSE=POS][TAM1=FUTURE][AGR=3SG]]
[PUNC=PERIOD]
```

After sending each word to the morphological generation system, the surface form of the sentence appears as follows:

Tosco ülkenin en büyük bağımsız rafinerisi olacak.

4 Results and Example Translations

For evaluating our prototype system, we have used 52 sentences (646 words) from a corpus of broadcast news captions [9]. Of these 52 sentences, the system was able to translate 44 sentences (85%) correctly and completely. 2 sentences (4%) had missing phrases because of problems with the mapping and sentence generation systems. 6 sentences (11%) could not be translated because of problems with the interlingua and the mapper. The reasons for the missing translations can be summarized as: (i) structural problems in the interlingua, such as incorrect prepositional phrase attachments; (ii) feature mismatches (values stored under wrong features); and (iii) mapper limitations (inability to implement certain mapping operations).

The following examples demonstrate the output of the system.

1. `(translate "The company says they have sealed the deal.")`
   `Şirket onların anlaşmayı imzaladıklarını söylüyor.`

2. `(translate "Tosco had sealed the deal in World War II.")`
   `Tosco II. dünya savaşında anlaşmayı imzalamıştı.`
   `Tosco II. dünya savaşındaki anlaşmayı imzalamıştı.`

Note that in the second example there are two valid translations in Turkish corresponding to the attachment ambiguities of the prepositional phrase in the English input. In the first Turkish sentence the prepositional phrase *in World War II* maps to a temporal adjunct, while in the second sentence, it maps to a relativizer noun phrase modifier.

5 Issues and Problems

In this section, we will discuss some issues related to the generation of Turkish text from an interlingua representation, and present how we have handled them. These issues can be categorized into three groups, according to their origin.

While there are challenges to be worked out where the source and target languages differ greatly in their means of realization for the same unit of meaning, our experience has been that the interlingua approach is an advantage when integrating software modules for languages that differ in their grammatical structure. For this reason we feel that our results should encourage others working on dissimilar language pairs to consider the interlingua approach. Many of the general issues listed below (e.g., tense differences, argument mappings, verb mapping, lexical selection) are not specific to English and Turkish, and the approaches described herein can be adapted for use with other language pairs.

5.1 Issues Related to the Differences between English and Turkish

Tense differences. There are some differences between tenses in Turkish and English. Some English tenses do not have exact Turkish counterparts, and vice-

versa. An example is the *narrative past tense* of Turkish, which is used when the speaker is talking about a past event, which she has not witnessed herself. Similarly, the past perfect and present perfect tenses of English do not have one-to-one correspondences in Turkish, hence they are mapped to the closest possible Turkish tenses.

Argument mappings. The KANT interlingua categorizes verbs according to their argument classes [6], which facilitates the mapping process. For example, in the case of a verb of argument class AGENT+THEME, agent maps to subject and the theme usually maps to accusative object. But, there are certain verbs that belong to the AGENT+THEME argument class, whose theme maps to a dative object in Turkish. For example despite the fact that the verbs 'break' and 'cause' belong to the AGENT+THEME arguments class, but 'break' in Turkish subcategorizes for an accusative object, whereas 'cause' subcategorizes for a dative object.

(3)a. Kedi vazo-yu kır-dı.
`Cat vase-ACC break-PAST-3SG`
'The cat broke the vase.'

b. Kedi kaza-ya sebep oldu.
`Cat accident-DAT cause-PAST-3SG`
'The cat caused an accident.'

Since such subcategorization information cannot be deduced from the interlingua, we introduced a SUBCAT feature. This feature stores the subcategorization information of the verb in the interlingua and is used during mapping. We map the arguments according to this feature, in addition to the argument class of the verb and the voice of the sentence.

Prepositional phrase attachments. Because of the prepositional phrase attachments, some sentences are inherently ambiguous in English. For example, for the English sentence "I saw the girl at home." it is possible to have two different interlingua representations. But, these two interlingua representations will map to different translations in Turkish.

(4)a. Ev-de kız-ı gör-dü-m.
`ev-LOC girl-ACC see-PAST-1SG`
'[I] [saw] [the girl] [at home].'

b. Ev-de-ki kız-ı gör-dü-m.
`home-LOC-REL girl-ACC see-PAST-1SG`
'[I] [saw] [the girl [at home]].'

Since the parser produces both interlingua representations, our system produces two surface forms for such sentences.

Additionally, certain prepositional phrases map to different structures in Turkish. A typical example is the preposition 'for'. If it is used for stating a

price or a beneficiary, it maps to a dative object in Turkish, otherwise it maps to a Turkish postpositional phrase, whose postposition is 'için'.

(5)a. Kitabı 7 dolara satın aldı.
`book-ACC 7 dollar-DAT buy-PAST-3SG`
'(He) bought the book for 7 dollars.'

b. Kitabı Ali'ye satın aldı.
`book-ACC Ali-DAT buy-PAST-3SG`
'(He) bought the book for Ali.'

c. O şirket için önemliydi.
`He company for important-PAST-3SG`
'He was important for the company.'

We generate the correct sentence by certain semantic checks. It is important to note that it is not always possible to preserve source text ambiguity when mapping to Turkish, because both source meanings cannot be indicated by a single output structure. For this reason disambiguation via semantic restrictions becomes crucial when mapping from English to Turkish.

Verb mappings. There are some verbs whose argument classes depend on their sentential context. For example, the verb 'finish' belongs to argument class THEME/AGENT+THEME in English. In the following sentence, it belongs to the THEME argument class which maps to *bit* in Turkish:

(6) The film finished.
Film bit-ti.
Film finish-PAST-3SG

On the other hand 'finish' belongs to the AGENT+THEME argument class in the sentence:

(7) He finished the school.
O okul-u bit-ir-di.
He school-ACC finish-CAUS-PAST-3SG

As can be seen from the glosses, these verbs have different surface realizations in Turkish. For example, in sentence (7), the verb has a CAUSATIVE marker, which is absent in the sentence (6), although the verbs have the same form in English. This is the case for all of the verbs in this argument class. In order to handle such cases, we make a test in the lexicon and add the causative marker if a verb has an AGENT+THEME argument class.

Lexical selection. Lexical selection is also an important issue for an MT system. As exemplified by (8a), the verb "say" is mapped to Turkish verb "de", while in (8b) it is mapped to the verb "söyle". The rationale for this selection is as follows: if there is a THEME feature in interlingua representation of the sentence, "say" maps to the verb "de", otherwise if there is a complement, it maps to the verb "söyle".

(8)a. John Mary'e olmaz dedi.
`John Mary-DAT no say-PAST-3SG`
'John said no to Mary.'
b. John geldiğini söyledi.
`John come-INF-ACC say-PAST-3SG`
'John said he came.'

Demonstrative pronoun mappings. Two demonstrative pronouns are used in English to denote singular concepts: 'this' and 'that', used for showing near and far objects, respectively. However Turkish employs three demonstrative pronouns for this purpose: 'bu', 'şu', and 'o', used for showing near, far, and very far objects, respectively. 'This' always maps to 'bu', but 'that' sometimes maps to 'şu', and sometimes to 'o', depending on the context. Since the distance information cannot be deduced as either "far" or "very far" from English, 'that' is always mapped to 'o' in this system.

5.2 Issues Related to the Interlingua

Anaphora resolution. The current KANT parser does not resolve anaphora. This resolution can be critical for Turkish. For instance for the sentence *'Ed read his book.'*, if the writer or owner of the book is Ed himself, the Turkish sentence that must be generated is:

(9) Ed kitab-ı-nı oku-du.
`Ed book-P3SG-ACC read-PAST-3SG`

Otherwise (i.e. the book belongs to or is written by another person), there must be an explicit pronoun with a genitive marker:

(10) Ed o-nun kitab-ı-nı oku-du.
`Ed he-GEN book-P3SG-ACC read-PAST-3SG`

5.3 Issues Related to the Generation and Mapping Systems

Word order variations. The mapping system does not currently produce an information structure (e.g., marking constituents as *topic, focus* or *background*). Such information when available is used by the generator to handle word order variations. So, currently all sentences are produced in the ical order (SOV) in Turkish. The information structure of a sentence can be obtained using syntactic clues in the source language in machine translation [1, 11], or using algorithms that determine the topic and focus of the target language sentences using Centering Theory [12], and given versus new information [3].

Domain differences. The sentence generation system was originally developed for a machine translation system in another domain [14]. Missing parts, like detailed treatment of numbers, were added during the development of the mapping system.

Mapper limitations. Features belonging to the same category are stored in the same slot in the interlingua, using a `:multiple` flag. The problem is that features belonging to the same category in the interlingua may map to different categories in Turkish. Currently, the mapper does not support the operation of extracting individual features under the `:multiple` flag.

6 Future Work

We have presented a system which generates Turkish sentences from interlingua representations. This work is important because it demonstrates the feasibility of rapidly combining independent systems developed at different locations, using interlingua as an intermediary representation. With the implementation of a Turkish mapping component, we were able to construct a prototype English–Turkish machine translation system in about two months.

The coverage, accuracy, and fluency of this machine translation system can further be extended, by adding new and more detailed mapping rules. For the example set of 52 sentences, the output quality of this system is comparable to the output quality of the KANT machine translation system [8] in large-scale domains. To achieve the same output results on a large-scale English–Turkish corpus, significant work must be undertaken to extend the lexicon and mapping rules.

7 Acknowledgments

This research has been supported in part by a NATO Science for Stability Grant TU–LANGUAGE. The authors would like to thank Robert Igo and Krzysztof Czuba for their help with the Turkish lexicon and mapper. We also thank Zelal Güngördü for extensive comments on earlier versions of this manuscript which significantly improved its presentation.

References

1. Eva Hajičová, Petr Sgall, and Hana Skoumalová. Identifying Topic and Focus by an Automatic Procedure. *Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics*, 1993.
2. Dilek Zeynep Hakkani and Kemal Oflazer. Tactical Generation in a Free Constituent Order Language. *Journal of Natural Language Engineering*, 4(2), June 1998.
3. Beryl Hoffman. Translating into Free Word Order Languages. In *Proceedings of COLING'96*, Copenhagen, Denmark, 1996.
4. John R. Leavitt. *KANT Mapper Specification.* Carnegie Mellon University–Center for Machine Translation, 1993.
5. G. L. Lewis. *Turkish Grammar.* Oxford University Press, 1991.
6. Teruko Mitamura and Eric H. Nyberg. Hierarchical Lexical Structure and Interpretive Mapping in Machine Translation. In *Proceedings of COLING-92*, 1992.
7. Eric H. Nyberg and Teruko Mitamura. The KANT System; Fast, Accurate, High-Quality Translation in Practical Domains. In *Proceedings of COLING'92*, Nantes, France, July 1992.
8. Eric H. Nyberg, T. Mitamura, and J. G. Carbonell. Evaluation Metrics for Knowledge-Based Machine Translation. In *Proceedings of COLING'94*, 1994.
9. Eric H. Nyberg and Teruko Mitamura. A Real-Time MT System for Translating Broadcast Captions. In *Proceedings of the Machine Translation Summit VI*, pages 51 – 57, 1997.
10. Kemal Oflazer. Two-level Description of Turkish Morphology. In *Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics*, April 1993. A full version appears in *Literary and Linguistic Computing*, Vol.9 No.2, 1994.
11. Ralf Steinberger. Treating Free Word Order in Machine Translation. *COLING, Kyoto, Japan*, 1994.
12. Malgorzata E. Styś and Stefan S. Zemke. Incorporating Discourse Aspects in English –Polish MT: Towards Robust Implementation. *Recent Advances in NLP*, 1995.
13. Masaru Tomita and Eric H. Nyberg. Generation Kit and Transformation Kit, Version 3.2, User's Manual. Technical report, Carnegie Mellon University–Center for Machine Translation, 1988.
14. C. K. Turhan. An English to Turkish Machine Translation System Using Structural Mapping. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, pages 320–323, 1997.
15. E. Vallduví. The Dynamics of Information Packaging. In Elizabeth Engdahl, editor, *Integrating Information Structure into Constraint-based and Categorial Approaches*, Esprit Basic Research Project 6852, DYANA-2, Report R1.3B. September 1994.

This article was processed using the LaTeX macro package with LLNCS style