Machine Translation between Turkic Languages

A. Cüneyd TANTUĞ
Istanbul Technical University
Istanbul, Turkey
tantug@itu.edu.tr

Eşref ADALI
Istanbul Technical University
Istanbul, Turkey
adali@itu.edu.tr

Kemal OFLAZER
Sabanci University
Istanbul, Turkey
oflazer@sabanciuniv.edu

Abstract

We present an approach to MT between Turkic languages and present results from an implementation of a MT system from Turkmen to Turkish. Our approach relies on ambiguous lexical and morphological transfer augmented with target side rule-based repairs and rescoring with statistical language models.

1 Introduction

Machine translation is certainly one of the toughest problems in natural language processing. It is generally accepted however that machine translation between close or related languages is simpler than full-fledged translation between languages that differ substantially in morphological and syntactic structure. In this paper, we present a machine translation system from Turkmen to Turkish, both of which belong to the Turkic language family. Turkic languages essentially exhibit the same characteristics at the morphological and syntactic levels. However, except for a few pairs, the languages are not mutually intelligible owing to substantial divergences in their lexicons possibly due to different regional and historical influences. Such divergences at the lexical level along with many but minor divergences at morphological and syntactic levels make the translation problem rather non-trivial. Our approach is based on essentially morphological processing, and direct lexical and morphological transfer, augmented with substantial multi-word processing on the source language side and statistical processing on the target side where data for statistical language modelling is more readily available.

2 Related Work

Studies on machine translation between close languages are generally concentrated around certain Slavic languages (e.g., Czech→Slovak, Czech→Polish, Czech→Lithuanian (Hajic et al., 2003)) and languages spoken in the Iberian Peninsula (e.g., Spanish↔Catalan (Canals et al., 2000), Spanish↔Galician (Corbi-Bellot et al., 2003) and Spanish↔Portugese (Garrido-Alenda et al., 2003). Most of these implementations use similar modules: a morphological analyzer, a part-of-speech tagger, a bilingual transfer dictionary and a morphological generator. Except for the Czech→Lithuanian system which uses a shallow parser, syntactic parsing is not necessary in most cases because of the similarities in word orders. Also, the lexical semantic ambiguity is usually preserved so, none of these systems has any module for handling the lexical ambiguity. For Turkic languages, Hamzaoğlu (1993) has developed a system from Turkish to Azerbaijani, and Altıntaş (2000) has developed a system from Turkish to Crimean Tatar.

3 Turkic Languages

Turkic languages, spoken by more than 180 million people, constitutes subfamily of Ural-Altaic languages and includes languages like Turkish, Azerbaijani, Turkmen, Uzbek, Kyrghyz, Kazakh, Tatar, Uyghur and many more. All Turkic languages have very productive inflectional and derivational agglutinative morphology. For example the Turkish word *evlerimizden* has three inflectional morphemes attached to a noun root *ev* (house), for the plural form with second person plural possessive agreement and ablative case:

Proceedings of the ACL 2007 Demo and Poster Sessions, pages 189–192,
Prague, June 2007.

```
evlerimizden    (from our houses)
ev+ler+imiz+den
ev+Noun+A3pl+P1sg+Abl
```

All Turkic languages exhibit SOV constituent order but depending on discourse requirements, constituents can be in any order without any substantial formal constraints. Syntactic structures between Turkic languages are more or less parallel though there are interesting divergences due to mismatches in multi-word or idiomatic constructions.

4 Approach

Our approach is based on a direct morphological transfer with some local multi-word processing on the source language side, and statistical disambiguation on the target language side. The main steps of our model are:

1. Source Language (SL) Morphological Analysis
2. SL Morphological Disambiguation
3. Multi-Word Unit (MWU) Recognizer
4. Morphological Transfer
5. Root Word Transfer
6. Statistical Disambiguation and Rescoring (SLM)
7. Sentence Level Rules (SLR)
8. Target Language (TL) Morphological Generator

Steps other than 3, 6 and 7 are the minimum requirements for a direct morphological translation model (henceforth, the baseline system). The MWU Recognizer, SLM and SLR modules are additional modules for the baseline system to improve the translation quality.

Source language morphological analysis may produce multiple interpretation of a source word, and usually, depending on the ambiguities brought about by multiple possible segmentations into root and suffixes, there may be different root words of possibly different parts-of-speech for the same word form. Furthermore, each root word thus produced may map to multiple target root words due to word sense ambiguity. Hence, among all possible sentences that can be generated with these ambiguities, the most probable one is selected by using various types of SLMs that are trained on target language corpora annotated with disambiguated roots and morphological features.

MWU processing in Turkic languages involves more than the usual lexicalized collocations and involves detection of mostly unlexicalized intra-word morphological patterns (Oflazer et al., 2004). Source MWUs are recognized and marked during source analysis and the root word transfer module maps these either to target MWU patterns, or directly translates when there is a divergence.

Morphological transfer is implemented by a set of rules hand-crafted using the contrastive knowledge between the selected language pair.

Although the syntactic structures are very similar between Turkic languages, there are quite many minor situations where target morphological features marking features such as subject-verb agreement have to be recovered when such features are not present in the source. Furthermore, occasionally certain phrases have to be rearranged. Finally, a morphological generator produces the surface forms of the lexical forms in the sentence.

5 Turkmen to Turkish MT System

The first implementation of our approach is from Turkmen to Turkish. A general diagram of our MT system is presented in Figure 1. The morphological analysis on the Turkmen side is performed by a two-level morphological analyzer developed using Xerox finite state tools (Tantuğ et al., 2006). It takes a Turkmen word and produces all possible morphological interpretations of that word. A simple experiment on our test set indicates that the average Turkmen word gets about 1.55 analyses. The multi-word recognition module operates on the output of the morphological analyzer and wherever applicable, combines analyses of multiple tokens into a new analysis with appropriate morphological features. One side effect of multi-word processing is a small reduction in morphological ambiguity, as when such units are combined, the remaining morphological interpretations for these tokens are deleted.

The actual transfer is carried out by transferring the morphological structures and word roots from the source language to the target language maintaining any ambiguity in the process. These are implemented with finite state transducers that are compiled from replace rules written in the Xerox regular expression language.[1] A very simple example of this transfer is shown in Figure 2.[2]

[1]The current implementation employs 28 replace rules for morphological feature transfer and 19 rules for sentence level processing.

[2]+Pos:Positive polarity, +A3sg: 3^{rd} person singular agreement, +Inf1,+Inf2: infinitive markers, +P3sg, +Pnon: possessive agreement markers, +Nom,+Acc: Nominative and ac-

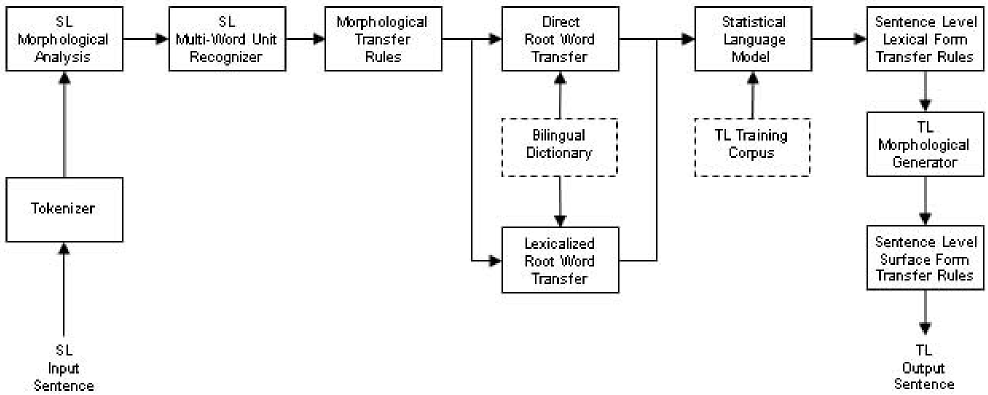

Figure 1: Main blocks of the translation system

`ösmegi`
↓
Source Morphological Analysis
↓
`ös+Verb+Pos^DB+Noun+Inf1+A3sg+P3sg+Nom`
`ös+Verb+Pos^DB+Noun+Inf1+A3sg+Pnon+Acc`
↓
Source-to-Target Morphological Feature Transfer
↓
`ös+Verb+Pos^DB+Noun+Inf2+A3sg+P3sg+Nom`
`ös+Verb+Pos^DB+Noun+Inf2+A3sg+Pnon+Acc`
↓
Source-to-Target Root word Transfer
↓
`ilerle+Verb+Pos^DB+Noun+Inf2+A3sg+P3sg+Nom`
`ilerle+Verb+Pos^DB+Noun+Inf2+A3sg+Pnon+Acc`
`büyü+Verb+Pos^DB+Noun+Inf2+A3sg+P3sg+Nom`
`büyü+Verb+Pos^DB+Noun+Inf2+A3sg+Pnon+Acc`
↓
Target Morphological Generation
↓
`ilerlemesi` (the progress of (something))
`ilerlemeyi` (the progress (as direct object))
`büyümesi` (the growth of (something))
`büyümeyi` (the growth (as direct object))

Figure 2: Word transfer

In this example, once the morphological analysis is produced, first we do a morphological feature transfer mapping. In this case, the only interesting mapping is the change of the infinitive marker. The source root verb is then ambiguously mapped to two verbs on the Turkish side. Finally, the Turkish surface form is generated by the morphological generator. Note that all the morphological processing details such as vowel harmony resolution (a morphographemic process common to all Turkic languages though not in identical ways) are localized to morphological generation.

Root word transfer is also based on a large transducer compiled from bilingual dictionaries which contain many-to-many mappings. During mapping this transducer takes into account the source root word POS.[3] In some rare cases, mapping the word root is not sufficient to generate a legal Turkish lexical structure, as sometimes a required feature on the target side may not be explicitly available on the source word to generate a proper word. In order to produce the correct mapping in such cases, some additional lexicalized rules look at a wider context and infer any needed features.

While the output of morphological feature transfer module is usually unambiguous, ambiguity arises during the root word transfer phase. We attempt to resolve this ambiguity on the target language side using statistical language models. This however presents additional difficulties as any statistical language model for Turkish (and possibly other Turkic languages) which is built by using the surface forms suffers from data sparsity problems. This is due to agglutinative morphology whereby a root word may give rise to too many inflected forms (about a hundred inflected forms for nouns and much more for verbs; when productive derivations are considered these numbers grow substantially!). Therefore, instead of building statistical language models on full word forms, we work with morphologically analyzed and disambiguated target language corpora. For example, we use a language model that is only based on the (disambiguated) root words to disambiguate ambiguous root words that arise from root

cusative case markers.

[3]Statistics on the test set indicate that on the average each source language root word maps to about 2 target language root words.

word transfer. We also employ a language model which is trained on the last set of inflectional features of morphological parses (hence does not involve any root words.)

Although word-by-word translation can produce reasonably high quality translations, but in many cases, it is also the source of many translation errors. To alleviate the shortcomings of the word-by-word translation approach, we resort to a series of rules that operate across the whole sentence. Such rules operate on the lexical and surface representation of the output sentence. For example, when the source language is missing a subject agreement marker on a verb, this feature can not be transferred to the target language and the target language generator will fail to generate the appropriate word. We use some simple heuristics that try to recover the agreement information from any overt pronominal subject in nominative case, and that failing, set the agreement to 3^{rd} person singular. Some sentence level rules require surface forms because this set of rules usually make orthographic changes affected by previous word forms. In the following example, suitable variants of the clitics *de* and *mi* must be selected so that vowel harmony with the previous token is preserved.

`o d`**`e`**` gördü m`**`i`**`?` → `o d`**`a`**` gördü m`**`ü`**`?`
`(did he see too?)`

A wide-coverage Turkish morphological analyzer (Oflazer, 1994) made available to be used in reverse direction to generate the surface forms of the translations.

6 Results and Evaluation

We have tracked the progress of our changes to our system using the BLEU metric (Papineni et al., 2004), though it has serious drawbacks for agglutinative and free constituent order languages.

The performance of the baseline system (all steps above, except 3, 6, and 7) and systems with additional modules are given in Table 1 for a set of 254 Turkmen sentences with 2 reference translations each. As seen in the table, each module contributes to the performance of the baseline system. Furthermore, a manual investigation of the outputs indicates that the actual quality of the translations is higher than the one indicated by the BLEU score.[4] The errors mostly stem from the statical language models not doing a good job at selecting the right root words and/or the right morphological features.

System	BLEU Score
Baseline	26.57
Baseline + MWU	28.45
Baseline + MWU + SLM	31.37
Baseline + MWU + SLM + SLR	33.34

Table 1: BLEU Scores

7 Conclusions

We have presented an MT system architecture between Turkic languages using morphological transfer coupled with target side language modelling and results from a Turkmen to Turkish system. The results are quite positive but there is quite some room for improvement. Our current work involves improving the quality of our current system as well as expanding this approach to Azerbaijani and Uyghur.

Acknowledgments

This work was partially supported by Project 106E048 funded by The Scientific and Technical Research Council of Turkey. Kemal Oflazer acknowledges the kind support of LTI at Carnegie Mellon University, where he was a sabbatical visitor during the academic year 2006 – 2007.

References

A. Cüneyd Tantuğ, Eşref Adalı, Kemal Oflazer. 2006. Computer Analysis of the Turkmen Language Morphology. *FinTAL, Lecture Notes in Computer Science*, 4139:186-193.

A. Garrido-Alenda et al. 2003. Shallow Parsing for Portuguese-Spanish Machine Translation. *in TASHA 2003: Workshop on Tagging and Shallow Processing of Portuguese*, Lisbon, Portugal.

A. M. Corbi-Bellot et al. 2005. An open-source shallow-transfer machine translation engine for the Romance languages of Spain. *in 10th EAMT conference "Practical applications of machine translation"*, Budapest, Hungary.

Jan Hajic, Petr Homola, Vladislav Kubon. 2003. A simple multilingual machine translation system. *MT Summit IX*.

İlker Hamzaoğlu. 1993. Machine translation from Turkish to other Turkic languages and an implementation for the Azeri language. *MSc Thesis, Bogazici University, Istanbul*.

Kemal Altıntaş. 2000. Turkish to Crimean Tatar Machine Translation System. *MSc Thesis, Bilkent University, Ankara*.

Kemal Oflazer. 1994. Two-level description of Turkish morphology. *Literary and Linguistic Computing*, 9(2).

Kemal Oflazer, Özlem Çetinoğlu, Bilge Say. 2004. Integrating Morphology with Multi-word Expression Processing in Turkish. *The ACL 2004 Workshop on Multiword Expressions: Integrating Processing*.

Kishore Papineni et al. 2002. BLEU : A Method for Automatic Evaluation of Machine Translation. *Association of Computational Linguistics, ACL'02*.

Raul Canals-Marote et al. 2000. interNOSTRUM: a Spanish-Catalan Machine Translation System. *Machine Translation Review*, 11:21-25.

[4]There are many translations which preserve the same meaning with the references but get low BLEU scores.