A MT System from Turkmen to Turkish Employing Finite State and Statistical Methods

A. Cüneyd TANTUĞ
Department of Computer Engineering
Istanbul Technical University
Maslak, 34469, Istanbul
Turkey
tantug@itu.edu.tr

Eşref ADALI
Department of Computer Engineering
Istanbul Technical University
Maslak, 34469, Istanbul
Turkey
adali@itu.edu.tr

Kemal OFLAZER
Faculty of Eng. and Natural Sciences
Sabanci University
Tuzla, 34956, Istanbul
Turkey
oflazer@sabanciuniv.edu

Abstract

In this work, we present a MT system from Turkmen to Turkish. Our system exploits the similarity of the languages by using a modified version of direct translation method. However, the complex inflectional and derivational morphology of the Turkic languages necessitate special treatment for word-by-word translation model. We also employ morphology-aware multi-word processing and statistical disambiguation processes in our system. We believe that this approach is valid for most of the Turkic languages and the architecture implemented using FSTs can be easily extended to those languages.

Introduction

Implementing a fully-automatic machine translation (MT) system capable of producing high-quality translations and handling unrestricted text, is still one of the most challenging tasks in natural language processing community. The more apart the source language (SL) and target language (TL) in terms of lexical, morphological and syntactical structures, the harder is the translation process (Nagao 1984; Jurafsky & Martin 2000). Recent advances in statistical machine translation has provide a new avenue to alleviate the complexities of MT but such systems rely on the availability of large amounts of parallel text which is not available for many language pairs. On the other hand, MT between close language pairs can be relatively easier and can still benefit from simple(r) paradigms in MT.

In this paper, we present a MT system from Turkmen language to Turkish. Both SL and TL are cognate Turkic languages sharing very similar linguistic features such as agglutinative morphology and word order. Our system is designed to translate using a direct translation motivated method augmented with a disambiguation post-processing stage based on statistical language models. The very productive inflectional and derivational morphology of Turkic languages however necessitate considerable modifications be made for not only conventional components of standard direct translation model but also in the application of statistical language modelling techniques.

We start with a summary of previous work on MT between related languages. Following that, we present brief information about Turkmen and Turkish, together with the common linguistic properties and divergences. We then describe then describe the details of our MT system, explain our evaluation methodology, and give resulting scores and sample translations. The last section is devoted to the conclusions and future directions.

Related Work

While MT has had a long history, work on MT between close language pairs is relatively recent. As far as we know, the first effort was the RUSLAN project which aimed at translating main-frame documents from Czech to Russian (Hajič 1987). Experience from this project was used in another MT project, Česilko, between two Slavonic languages, Czech and Slovak (Hajič *et al.* 2000). This work was extended to cover some other Slavonic languages like Polish and Lower Serbian (Dvořák *et al.* 2006) and Lithuanian (Hajič *et al.* 2003), which is actually a Baltic language.

Following the development of interNOSTRUM project translating between Catalan and Spanish (Canals-Marote *et al.* 2000), additional work for other Romance Language pairs emerged. A Portuguese-Spanish MT system was implemented in the same manner (Garrido-Alenda *et al.* 2003). As a result of these projects, an open source shallow-transfer MT engine for the Romance languages of Spain has been implemented and made available to the public (M.Corbi-Bellot *et al.* 2005).

All of the systems mentioned above have very similar word-by-word translation architectures that have four basic components: (1) a morphological analyzer, (2) a POS-Tagger, (3) a transfer module and (4) a morphological generator. As they operate on very close language pairs for which syntactical analysis stage is unnecessary during translation, a word-by-word translation is usually adequate (except for the Czech-Lithuanian case where a shallow parser was used). Even homonyms preserve their homonymy after translation, so one-to-one word mapping works fine for these language pairs.

Work on MT between Turkic languages is also relatively limited. Hamzaoğlu (1993) presented a lexicon based MT system from Turkish to Azerbaijani, which is probably the language closest to Turkish. A more recent work involves a Turkish to Crimean Tatar MT system which is able to generate ambiguous translations with a limited dictionary (Altıntaş & Çiçekli 2001).

Turkish and Turkmen Languages

As a subfamily of Ural-Altaic language family, Turkic languages comprise over 40 languages: Turkish, Azerbaijani, Uzbek, Turkmen, Kyrgyz, Kazakh, Uighur, Chagatai, Karagas, Tatar, Yakut, Chuvash being the more prominent ones[1]. Turkish is the largest Turkic language having more than 70 million native speakers while Turkmen language is used by approximately 11 million people. Turkmen language shares many common linguistic properties with Turkish to some extent. However, many divergences due to regional and historical reasons prevent the mutual intelligibility across these languages.

Morphology

Both Turkish and Turkmen language have very productive derivational and inflectional morphologies where suffixes are affixed to a root word or another suffix (Oflazer 1995). Here are two examples word formation in these languages:[2]

```
Turkish : sağlamlaştırdık      (we made it strong)

sağlam+Adj
   ^DB+Verb+Become
   ^DB+Verb+Caus+Pos+Past+A1pl


Turkmen : baglanyşyklydyr      ((it) is related to …)

baglanyşyk+Noun+A3sg+Pnon+Nom
       ^DB+Adj+With
       ^DB+Verb+Zero+Pres+Cop+A3sg
```

Although Turkmen and Turkish languages use different alphabets in their orthography, most of the morphophonetic rules are common. For example, both languages exhibit vowel harmony and consonant mutation rules.

Both of the languages include similar suffixes with same or very close semantics. However, divergences like different tense aspect moods or different subject-verb agreement properties are observed frequently. For instance, the *+makçı/+mekçi* Turkmen suffix[3] does not have counterpart in Turkish. Similarly, the definite future tense suffix attached to a Turkmen verb never accepts a person agreement marker after it; on the contrary, such a marker is a must in Turkish.

Turkmen morphotactics is very similar to Turkish, essentially for nominals. However, Turkmen verbal morphotactics differ from Turkish in many cases like denoting polarity and the order of causative, passive or reflexive suffixes.

The most problematic morphological issue for Turkic languages is ambiguity. Surface forms can be segmented in various ways and thus can result in different root words and/or different suffix combinations. An example of morphological ambiguity for the Turkish surface form *izin* is given below (Hakkani-Tür *et al.* 2002).

```
1. iz+Noun+A3sg+Pnon+Gen (trace/print)
2. iz+Noun+A3sg+P2sg+Nom (your trace/print)
3. izin+Noun+A3sg+Pnon+Nom (permission)
```

Similar kinds of ambiguities are also observed for Turkmen language.

Syntactic Structure

All Turkic languages are free constituent order languages; phrases can be arranged freely within the sentence based on discourse requirements. However the unmarked order is SOV. Morphological case markers of some constituents determine their grammatical role in the sentence.

From the point of view of syntactical structure, an almost one-to-one mapping can be observed between Turkish and Turkmen. However, word-by-word correspondence fails in many situations. Some Turkmen multi-word units (MWU) may be translated into Turkish as a single word. In many cases with adjective participles, it is inevitable to change the position of some morphemes among other words in the adjectival phrase. A sample alignment between a Turkmen sentence and a Turkish sentence is given below:

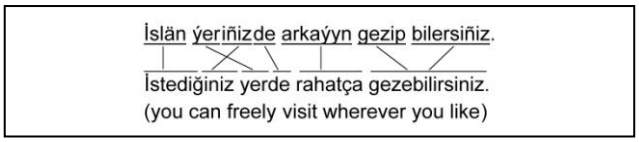

Figure 1-Alignment Example

In this alignment, one can readily see the replacement of Turkmen *+iñiz* morpheme[4] with its Turkish equivalent *+iniz*, and also the positional change of the morpheme from the noun to the participle adjectival form. Additionally, a typical instance of a case where SL MWU is aligned with a single TL word occurs in the end of the sentence. These and many other examples show that in spite of the syntactical similarities, word-by-word translation is not sufficient solely, and additional sentence level processing must be employed.

Lexicon

Since the origins of the Turkic languages are same, their lexicons share considerable amount of root words sometimes with only minor variations. Most of the variations are observed in orthography whereas spoken languages have more common patterns. Personal pronouns, date/time expressions, organ names, main color names, numbers are nearly same for all Turkic languages. As an example, the personal pronoun *ben* (I) in Turkish is preserved in most of the Turkic languages as its original state or with small phonetic variations (like *men* in Turkmen).

[1] http://ww.sil.org, 2007

[2] ***^DB*** denotes the derivation boundaries. Please refer to the appendix section for the glosses of other morphological markers

[3] Literal meaning : "***planning or thinking to do sth.***"

[4] 2nd person plural possesive agreement suffix

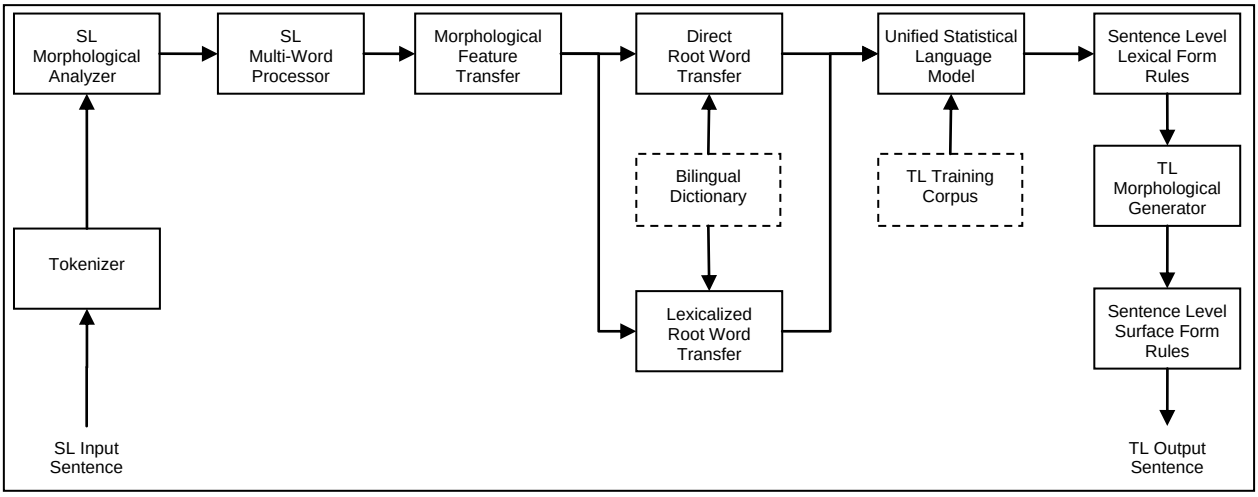

Figure 2 –Block Diagram of the System

The Translation System

We have designed and implemented a MT system from Turkmen to Turkish. Our system relies on morphological and lexical transfer with all possible ambiguities, and disambiguation by unified TL statistical language modelling. We have opted for disambiguating after conveying the SL side morphological ambiguities to the TL, because Turkmen is still a resource-poor language for statistical language modelling purposes. In order to alleviate the shortcomings of word-by-word transfer model, a morphology-aware MWU recognizer and a sentence level processing module are incorporated into the base system. The main blocks of the system are depicted in Figure 2. Except statistical disambiguation stage, all of the system components are designed in a finite state framework.[5]

The proposed method and implemented architecture can be used for MT between not only Turkmen to Turkish but also for other Turkic language pairs.

SL Morphological Analyser

By using Xerox Finite State tools (Karttunen *et al.* 1997), we have developed a two-level Turkmen morphological analyzer (Tantuğ *et al.* 2006).[6] It is noteworthy that the outputs of this morphological analyzer are ambiguous; the average number of outputs for each input word is about 1.55 analyses. This morphological analyzer is carefully designed to maintain the maximum intersection with the Turkish analyzer in hand, so that the number of morphological feature transfer rules can be kept minimal.

Multi-Word Processing

MWU processing in agglutinative languages is a quite difficult task owing to unlimited number of surface forms. Quick look-up lists for MWUs are useless for Turkish and Turkmen since the components of MWUs can suffer a derivational and/or inflectional process. In order to recognize MWUs, we used a very similar approach of the one described in (Oflazer *et al.* 2004) and classified SL MWUs into 3 separate classes:

1. Lexicalized MWUs
2. Semi-Lexicalized MWUs
3. Non-Lexicalized MWUs

A list of lexicalized MWUs is sufficient for recognition since all of the components in this type of collocations are fixed. However, semi-lexicalized collocations can undergo any kind of morphological process. For example, Turkish equivalents of Turkmen noun *gürüm-jürüm* and verb *bol-* are *gizli* (hidden) and *ol-* (to be), respectively. On the other hand, their combination *"gürüm-jürüm bol-"* should be translated as *kaybol-* (to lose). But, the phrase *"gürüm-jürüm bol-"* occurs usually in inflected and/or derived forms along the running text:

```
gürüm-jürüm bolypdyrsyñ      kaybolmuşsundur
                             (you must have been lost)
gürüm-jürüm boldy            kayboldu
                             ((he) is lost)
```

The MWU recognizer module uses both surface forms and morphological analyses to detect any possible MWU sequence, and if a match is found, it combines this sequence into a new parse:

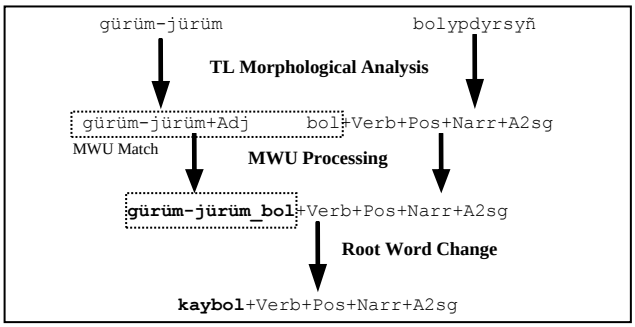

Figure 3 – MWU Processing Example

[5] Though conceivably we could have employed a weighted finite state toolkit to implement all components in a single-framework. But such toolkits do not provide the flexibility that we needed for the transfer module.

[6] The root word coverage of this morphological analyzer is relatively limited, we are currently working on extending it by adding new root words

In non-lexicalized type of MWUs, morphosyntactic patterns are used to construct the collocation. Some modal formations are expressed by the combination of two or more words, one or both of which is a verb. For instance, when placed after another verb ending *+yp/+ip*, the verb *ber-* (to give (to)) indicates that an action is performed for the benefit of someone else, while the same formation with *al-* (to take (from)) indicates an action performed for oneself:

```
okap berdi        (did read for sb.)

oka+Verb+Pos^DB+Adverb+AfterDoingSo ber+Verb+Pos+Past+A3sg

oku+Verb+Pos^DB+Verb+Hastily+Past+A3sg

okuyuverdi
```

SL root word and morphological features which are not enclosed by the upper dotted box, are variable parts of this type of collocation. While combining these two words, MWU recognizer also transforms the inside of the box into an appropriate TL structure.

Morphological Feature Transfer

This block includes morphological feature transfer rules realized in Xerox regular expression language. A total of 28 hand-crafted rules are prepared using the contrastive knowledge between languages. In the example rule below, the Turkmen aspect mood which has a meaning like "*intention* or *thinking/planning to do sth*" represented by the morphological marker *+Inten*, is transferred to Turkish as two separate words. An application of this rule with root word mapping is shown in Figure 4.

```
define ChangeInten "+Inten+Anon" ->
          "^DB+Noun+Inf1+A3sg+Pnon+Nom iste+Verb+Pos+Prog1+A3sg"
```

Root Word Transfer

This module takes all possible morphological analyses of input SL words and replaces the root of the parse with one or multiple TL word(s) selected from the bilingual root word transfer dictionary. Therefore, one-to-many mappings taking place in this process causes another type of ambiguity, namely lexical ambiguity. The part-of-speech of the root word is taken into account when performing this mapping, so that spurious mappings based on just the written form can be eliminated. The following example root transfer rules contain two different Turkish words, *gri (*gray) as adjective sense and *sil* (erase) as verb sense of Turkmen word *boz.* Please also note that the verb *geple* has two corresponding Turkish entries, *söyle* (say) and *konuş* (talk), producing ambiguous outputs, whereas in the former case, ambiguity is resolved by the help of POS information.

```
define AdjDict    "gri" <- "boz" \/ _ "+Adj" .o.
                  ...;

define VerbDict   "sil" <- "boz" \/ _ "+Verb" .o.
                  "söyle" <- "geple" \/ _ "+Verb" .o.
                  "konuş" <- "geple" \/ _ "+Verb" .o.
                  ...;
```

A sample of root word and morphological transfer process is given in Figure 4:

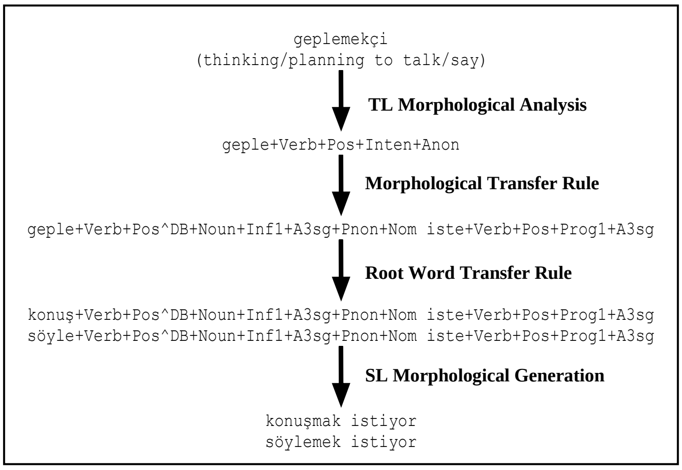

Figure 4 – Application of transfer rules

For some words, a simple replacement of the root word cannot produce a legal word form in the generation stage. This stems usually from the productive structure of the Turkic languages. For sake of clarity, consider the following Turkmen word *ulumsylyk* (vanity) and its morphological analysis.

```
ulumsylyk        ulumsy+Adj^DB+Noun+Ness+A3sg+Pnon+Nom
```

In the bilingual dictionary, *kibirli* (arrogant) is the corresponding Turkish root word for Turkmen root word *ulumsy*. However, a direct replacement of Turkmen root with its Turkish counterpart causes a failure in generation stage, because *kibirli* is not a legal root word in Turkish. In fact, *kibirli* is actually derived from the original root *kibir* (arrogance) with the suffix *+li* (with). So, to produce the right word form, a special lexicalized rule is required which replaces the *ulumsy+Adj* structure with the proper morphological representation of *kibirli*:

```
"kibir+Noun+A3sg+Pnon+Nom^DB+Adj+With" <- "ulumsy+Adj"
```

This type of lexicalized rules are not necessary for the SL words which are derived by the suffixes that have direct equivalents in TL with same semantics, such as *+lyk* (Turkmen) and *+lık* (Turkish) suffix pair.

In some rare cases, mapping the root and available morphological features is not sufficient to generate a legal Turkish lexical structure as sometimes some required feature on the target side may not be explicitly available on the source word. For such a case, we use rules that look at much wide context, mostly using additional heuristics to infer such features.

Statistical Disambiguation

To resolve both source side morphological ambiguities and lexical transfer ambiguities, we employ statistical language models (LM) on the TL side. A LM is normally generated by using surface forms; but this causes serious data sparseness problems for Turkic languages due to the agglutinative structure as the vocabulary size is quite

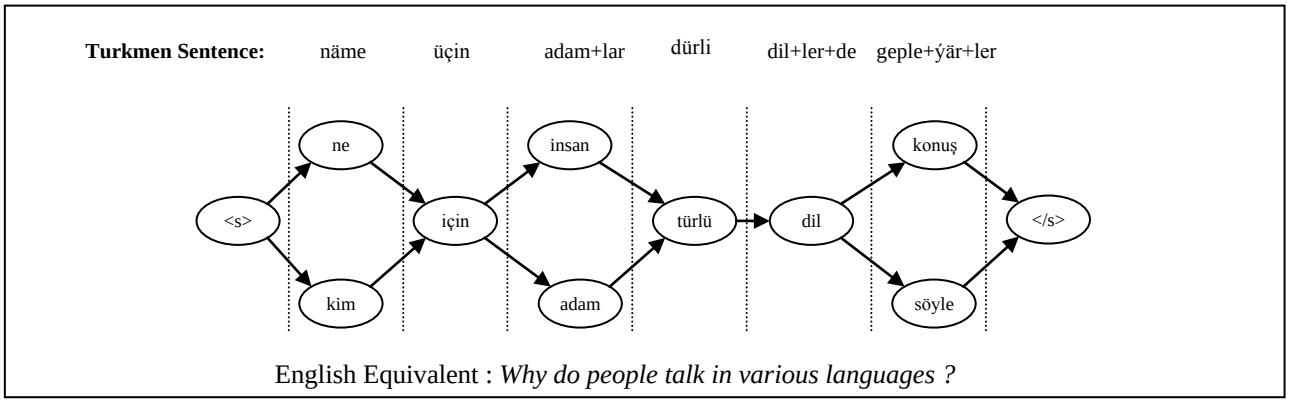

Figure 5 - A sample decoding of the candidate translations by using language models

large. Instead of building a single LM to model the full word forms, we have employed a unified language modelling concept where different LMs are utilized to model different parts of the language. As a first step, TL training corpora is morphologically analyzed and disambiguated to build various types of LMs. For example, one type of LM which is trained on only disambiguated root words can play an effective role in solving lexical ambiguity problems. In Figure 5, we show roots in Turkmen sentence with its Turkish root translations. The transition probabilities come from the LM probabilities. The most probable path (most probable translation) is found by the Viterbi algorithm. Table 1 shows the decoding of the example sentence in Figure 5, where the bold sentence indicates the right translation. The bi-gram LM achieves to resolve the ambiguities so that the right translation gained the first rank is selected as the output.

LM Order	Rank	Most Probable Sentences
Unigram	1	ne için insanlar türlü dillerde söylüyorlar
	2	**ne için insanlar türlü dillerde konuşuyorlar**
	3	ne için adamlar türlü dillerde söylüyorlar
Bigram	1	**ne için insanlar türlü dillerde konuşuyorlar**
	2	ne için adamlar türlü dillerde konuşuyorlar
	3	kim için insanlar türlü dillerde konuşuyorlar
Trigram	1	**ne için insanlar türlü dillerde konuşuyorlar**
	2	kim için insanlar türlü dillerde konuşuyorlar
	3	ne için adamlar türlü dillerde konuşuyorlar

Table 1 –Viterbi results of the sample sentence

Similarly, SL morphological ambiguities can be resolved by LMs trained on other morphological features (for instance, the last set of inflectional groups in the analyses or full morphological features except the root word).

Input of this statistical processing module is an ordered bag of all possible translations of the input sentence, including all kinds of ambiguities. LM based disambiguation module tries to find the most suitable sequence that have the highest probability value by constructing a HMM and running Viterbi algorithm on it.

Sentence Level Rules

These rules implement some modifications to overcome the problems of word-by-word transfer paradigm. These FST-based rules do some sentence level work such as finding adjective phrases and making morpheme arrangements within the phrase. Phrases are not determined by a parser since we do not need a full parse. Instead, we employ some chunking rules just for finding certain phrase patterns. In the following example, a morpheme of an adjective phrase discovered in the original sentence, is replaced and relocated to its expected grammatical place in Turkish.

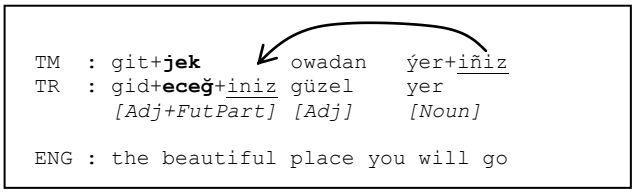

Figure 6 – Sentence level processing of an adjective phrase

Certainly, these kinds of rules require the morphological representation of the sentence. However, a small set of sentence level rules require the surface realizations of the words. Mostly, this set of rules deal with orthographic changes caused by the phonetic interactions between words. Note that the set of rules operating on surface forms must be performed after the TL morphological generation stage.

Morphological Generation

Morphological representation of the translated TL sentence is transformed into the final representation by generating surface forms of the words using a TL morphological generator. We have used a well-known wide-coverage Turkish morphological analyzer (Oflazer 1995) in reverse (generation) direction to synthesize resulting word forms. Except for very few cases, each Turkish morphological parse maps to only one surface form, hence no ambiguity arises during this generation phase.

Transliteration

Although it is not shown in the block diagram, a transliteration module functions in failure cases (i.e., morphological analysis or dictionary lookup failures) by exploiting the root word similarities of the languages. Instead of producing no output due to failures, our system produces TL equivalent by transliteration which

Sample 1

Turkmen: Bir adam çölüñ içinde kompas tapypdyr .
Turkish: Bir insan çölün içinde pusula bulmuştur .
Reference: Bir adam çölün içinde pusula bulmuş .

Sample 2

Turkmen: Hamyr gyzgyn tamdyrda esli wagtlap durandan soñ, ondan tagamly bir zadyñ ysy çykyp ugrapdyr.
Turkish: Hamur kızgın tandırda epeyce süre durduktan sonra, ondan tatlı bir şeyin kokusu çıkmaya başlamıştır.
Reference: Hamur sıcak tandırda epeyce durduktan sonra ondan lezzetli bir şeyin kokusu çıkmaya başlamış.

Sample 3

Turkmen: Sebäbi meniñ içimde goşa ýumruk ýaly gyzyl bardy .
Turkish: Nedeni benim içimde çift yumruk gibi altın vardı .
Reference: Çünkü benim içimde çifte yumruk kadar altın vardı .

Figure 7– Sample output translations from our Turkmen to Turkish MT system

sometimes increases the intelligibility but unfortunately this usually provides no improvement in our evaluation. In the example below, transliterated version of the Turkmen word *mylaýym* (mild) is more likely to let the native Turkish speaker to make an analogy with the original Turkish corresponding *mülayim*.

mylaýym —transliteration→ mılayım

Results and Evaluation

Outputs of the system are evaluated by the BLEU metric (Papineni *et al.* 2002). However, full-word form matching strategy of BLEU is inappropriate for agglutinative languages as a wrong suffix may cause a total mismatch even though the remaining part is completely right. Besides, the degree of variations in terms of word order is supposed to be compensated by multiple references, which does not always hold for free word order languages like Turkish. Nevertheless, we have used BLEU metric to measure the effects of our incremental changes on the system.

We get a BLEU score of 33.54 against a test set including 254 Turkmen sentences and 2 references per each sentence. In Figure 7, sample output translations are presented with their references. Since BLEU metric evaluates resulting translations very harshly, we employ an additional evaluation strategy by supplying only root words of the system outputs and references, and we get 39.31 as root BLEU (BLEU-r) score. This score denotes the success in choosing the right root words in target side. LMs that cannot find the right morphological features cause the difference between our BLEU and BLEU-r scores.

Current BLEU scores of our system outputs indicate that it can produce fairly high quality translations. Moreover, when manually evaluated, one can observe that some translations get very low scores even though they have very similar meanings with the references. This reveals that the actual quality of the translations is higher to the extent that BLEU scores indicate.

Error-Analysis

A detailed error analysis shows that erroneous situations mostly caused by wrong root word selections. In these cases, either the system chooses the false sense for the input word or the sequence with the highest probability contains a synonym of the reference words. Also, another distinct source of error is the group of Turkmen nouns ending with the vowel "*a*". Written Turkmen language has no discrimination between nominal case and dative case usage of these nouns whereas in the spoken language, last vowel *a* is stressed for the dative case. So, Turkmen morphological analyzer always produces the dative and nominal case analysis together for those words. As a result, a noun ending with the vowel *a* may be translated in a spurious dative or nominal case.

Conclusions and Future Work

In summary, we propose a methodology based on a direct translation model by exploiting similarities of the Turkic languages. Our approach includes word-by-word translation enhanced by additional morphological transferring phase, as well as morphology guided MWU processing. Since SL is a resource-poor language, we choose resolving both morphological and lexical ambiguities on the target side, by the help of unified statistical LMs specifically trained on agglutinative language corpora. To ease some drawbacks of word-by-word transfer strategy, we resort to apply a set of rules working on the sentence level.

Our next aim is to enlarge the size of our test set and references. Also, we are trying to extend the coverage of SL analyzer and bilingual dictionary, and to improve the quality of translations. As future work, we are currently working on other Turkic languages like Uighur and Azerbaijani.

Acknowledgements

This project is partly supported by The Scientific and Technical Research Council of Turkey under the contract no 106E048.

References

Altıntaş K. & Çiçekli İ. (2001) A Morphological Analyser for Crimean Tatar. In: Proceedings of the 10th Turkish Symposium on Artificial Intelligence and Neural Networks, TAINN, pp. 180-189, North Cyprus

Canals-Marote R., Esteve-Guillén A., Garrido-Alenda A., Guardiola--Savall M.I., Iturraspe-Bellver A., Montserrat-Buendia S., Pérez-Antón-Rojas P., Ortiz-Pina S., Pastor-Antón H. & Forcada M.L. (2000)

interNOSTRUM: a Spanish-Catalan Machine Translation System. Machine Translation Review, 11, 21-25

Dvořák B., Homola P. & Kuboň V. (2006) Exploiting similarity in the MT into a minority language. In: LREC-2006: Fifth International Conference on Language Resources and Evaluation. 5th SALTMIL Workshop on Minority Languages: "Strategies for developing machine translation for minority languages", Genoa, Italy

Garrido-Alenda A., Gilabert-Zarco P., Pérez-Ortiz J.A., Pertusa-Ibáñez A., Ramírez-Sánchez G., Sánchez-Martínez F., Scalco M.A. & Forcada M.L. (2003) Shallow Parsing for Portuguese-Spanish Machine Translation In: TASHA 2003: Workshop on Tagging and Shallow Processing of Portuguese, Lisbon, Portugal

Hajič J. (1987) RUSLAN - An MT System Between Closely Related Languages. In: Third Conference of the European Chapter of the Association for Computational Linguistics (EACL'87), Copenhagen, Denmark

Hajič J., Homola P. & Kuboň V. (2003) A simple multilingual machine translation system. In: MT Summit IX, New Orleans, USA

Hajič J., Hric J. & Kuboň V. (2000) Machine translation of very close languages. In: Proceedings of the sixth conference on Applied natural language processing pp. 7-12. Morgan Kaufmann Publishers Inc., Proceedings of the sixth conference on Applied natural language processing

Hakkani-Tür D.Z., Oflazer K. & Tür G. (2002) Statistical Morphological Disambiguation for Agglutinative Languages. Computers and the Humanities, 36, 381-410

Jurafsky D. & Martin J.H. (2000) Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Prentice Hall.

Karttunen L., Gaal T. & Kempe A. (1997) Xerox Finite-State Tool. In. XEROX Research Centre, Europe, Technical Report

M.Corbi-Bellot A., Forcada M.L., Ortíz-Rojas S., Pérez-Ortiz J.A., Ramírez-Sánchez G., Sánchez-Martínez F., Alegria I., Mayor A. & Sarasola K. (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

Nagao M. (1984) A Framework of a Mechanical Translation Between Japanese and English by Analogy Principle. In: Artificial and Human Intelligence (eds. by Banerji & Elithom), North-Holland

Oflazer K. (1995) Two-level Description of Turkish Morphology. Literary and Linguistic Computing, 9, 137-148

Oflazer K., Çetinoğlu Ö. & Say B. (2004) Integrating Morphology with Multi-word Expression Processing in Turkish. In: The ACL 2004 Workshop on Multiword Expressions:Integrating Processing, Barcelona, Spain

Papineni K., Roukos S., Ward T. & Zhu W.-J.J. (2002) BLEU : A Method for Automatic Evaluation of Machine Translation. In: Association of Computational Linguistics, ACL'02, Philadelphia, PA, USA

Tantuğ A.C., Adalı E. & Oflazer K. (2006) Computer Analysis of the Turkmen Language Morphology. In: FinTAL, Lecture Notes in Computer Science, pp. 186-193. Springer

Appendix – Glosses for Morphological Markers

+Anon : No agreement information
+A2sg : 2nd person singular agreement
+A3sg : 3rd person singular agreement
+A3pl : 3rd person plural agreement
+FutPart: Future participle
+Gen : Genitive case for nominals
+Narr : Narrative past tense (+mış/+miş/+muş/+müş)
+Nom : Nominative case for nominals
+Inf1 : Type 1 infinitive (+mak/+mek)
+Inten : Verbal mood of intention, thinking/planning to do sth. (+makçı/+mekçi)
+Pnon : No possessive agreement information
+P2pl : 2nd person plural possesive agreement
+Pos : Positive polarity
+Prog1 : Type 1 progressive tense (+iyor)