Using Finite State Transducers for Helping Foreign Language Learning

Hasan Kaya
Istanbul Technical University
Department of Computer Engineering
Istanbul, 34469, Turkey
kayahasa@itu.edu.tr

Gülşen Eryiğit
Istanbul Technical University
Department of Computer Engineering
Istanbul, 34469, Turkey
gulsenc@itu.edu.tr

Abstract

The interest and demand to foreign language learning are increased tremendously along with the globalization and freedom of movement in the world. Today, the technological developments allow the creation of supportive materials for foreign language learners. However, the language acquisition between languages with high typological differences still poses challenges for this area and the learning task it self. This paper introduces our preliminary study for building an educational application to help foreign language learning between Turkish and English. The paper presents the use of finite state technology for building a Turkish word synthesis system (which allows to choose word-related features among predefined grammatical affix categories such as tense, modality and polarity etc...) and a word-level translation system between the languages in focus. The developed system is observed to outperform the popular online translation systems for word-level translation in terms of grammatically correct outputs.

1 Introduction

The influence of mother tongue on foreign language learning is discussed in many linguistic and psychological studies (Hakuta et al., 2000; Hakuta, 1999; Durgunoglu and Hancin-Bhatt, 1992; Ringbom, 1987; Swan, 1997; Corder, 1983) in the literature. The typological differences between the mother tongue and the second language have an important role on the duration of learning process. In these studies, it is emphasized that one of the causes of frequently made mistakes in the second language is the rules learned from the first language. English and Turkish being languages from totally different language families compose a very representative and interesting language pair for this phenomena.

Turkish is an agglutinative language with a very rich morphological structure. Most of the syntactic information on the English side become morphological properties of a word on the Turkish side. In some cases, a single Turkish word may correspond to a full English sentence. This situation results in difficulties during language learning between this language pair and also in statistical machine translation (MT) systems. In daily life in Turkey, it is very common to come across with foreigners making mistakes in constructing Turkish words with invalid grammatical constructions (i.e. having difficulty to produce the correct morpheme order to form a valid Turkish word). Bisazza and Federico (2009), Yeniterzi and Oflazer (2010), El-Kahlout and Oflazer (2010) and Eyigöz et al. (2013) show the influence of using morphological clues in increasing the MT quality.

Finite state technology is proven to increase the efficiency in many rule-based NLP related tasks (Mohri, 1997; Roche and Schabes, 1997). Today, the availability of finite state transducer (FST) frameworks such as OpenFST (Allauzen et al., 2007), HFST (Lindén et al., 2009) and XFST (Beesley and Karttunen, 2003) makes possible to create FST applications very efficiently. In this paper, we present the results of our elementary studies on using finite state transducers to build supportive tools for foreign language learning; namely Turkish for English native speakers and English for Turkish native speakers. We compare our results with four popularly online translation systems: 1. Google[1], 2. Yandex[2], 3. Bing[3] and 4. Tureng[4]. The paper is organized as follows: Section 2 introduces the Turkish morphology, Section 3 the system architecture, Section 4 the learning use cases and Section 5 the conclusion and future work.

2 Turkish Morphology

As mentioned in the previous section, agglutinative language morphology has a high impact on the performance

[1] https://translate.google.com/
[2] https://ceviri.yandex.com.tr/
[3] http://www.bing.com/translator/
[4] http://tureng.com/search/translate

Proceedings of The 2nd Workshop on Natural Language Processing Techniques for Educational Applications, pages 94–98,
Beijing, China, July 31, 2015.

of translation process of a word. Turkish has a complex morphology and because of this reason, the usage of suffix concatenations at the end of a word lemma may cause the word to denote different meanings. Table 1 gives some translation examples from Turkish to English to show that the usage of dictionary/lexicon look-up systems are not suitable for word translation between Turkish and English due to unpredictable dictionary size. Foreign language learners have even difficulty to search for the meaning of a Turkish word from a Turkish dictionary since this task requires to firstly determine the lemma of that word. To give an example for this problem, word stem for "git" (*go*) can be written in different conjugated word forms such as: "gidiyorum", "gideceğim", "gidecek" etc. (up to nearly 50 variations) which refer to very different translations in the English side although the lemma of these words are the same. As a result, a Turkish word may be expressed as a single English word or a phrase or even a sentence as shown in Table 1.

Turkish	English
git	Go
gidiyorum	I am going
gideceğim	I'll go
gidecek	He will go
gittim	I went
gidebiliyor	He is able to go
gidebilmişlerdi	They had been able to go

Table 1: Translation of Turkish words

3 System Architecture

Eryiğit (2014) introduces a web service for morphological analysis and generation of Turkish. The provided analyzer is an updated version of the work presented in Şahin et al. (2013) and uses finite state technology for the analysis and generation purposes. In the provided interface the *surface word form* "gidiyorum" (*I'm going*) is analyzed as the *lexical form* "git+Verb+Pos+Prog1+A1sg" where "git" (*to go*) is the lemma of the word and the following tags hold for main parts-of-speech tag and additional inflectional features: "+Pos" for the positive marker, "Prog1" for the progressive tense, "A1sg" for the 1^{st} singular person. Similarly the same analysis given to the morphological generator produces the same input word. Inspired from this work, we develop a new finite state transducer transfer model and an English analyzer/generator which take the produced morphological analysis as input and produces its English counterpart. The system also works in reverse direction so that once an English input is given to the system, it transfers it to a Turkish lexical form and then uses the morphological generator to produce a valid Turkish word. Figure 1 draws the main flow of our system which we call "ITU Morphological Transfer module for English-Turkish language pairs"; ITUMorphTrans4ET in short from now on. The figure provides the intermediate stages for two given examples: "gidiyorum" (*I'm going*) and "gittim" (I went).

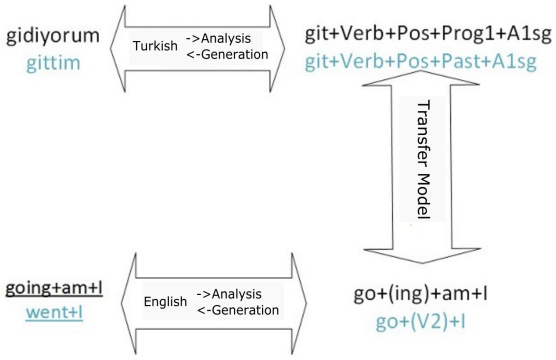

Figure 1: ITUMorphTrans4ET System Architechture

The transfer model is very similar to an FST morphological analyzer but instead of producing the relevant morphological tags for a given surface form, it produces a new lexical form (the English counterpart) of the input Turkish lexical form. To this end, it contains a bilingual lexicon for word lemmas and the transfer rules. An example from the transfer model FST is given below where each morphological tag (i.e. a suffix in the word surface form) representing person agreements are coded to produce two English words.

```
LEXICON Fin-ED-PC
+A1sg:+am+I #;
+A2sg:+are+you #;
+A3sg:+is+he/she/it #;
+A1pl:+are+we #;
+A2pl:+are+you #;
+A3pl:+are+they #;
```

The English output of the transfer model may be either some words or some tags to be further processed by the lexical post-processor. It is not necessary that all the Turkish tags produce an output; some of them are only required for determining the possible paths on the FST. This may be observed on the lexical forms of Turkish words in Figure 1. While the "+Prog1" tag is changed to an "+ing" tag, the "+Past" tag is changed to a "+V2" tag and the "+A2sg" tag is changed to the word "+he/she/it" in the transfer model's output, the "+Verb" or "+Noun" tags are only used for forwarding the process to possible FST paths in the coded rules.

The English analysis and generation FST converts between lexical and surface forms of English inputs. One should keep in mind that the English analyzer differing from the Turkish one also accepts inputs with multiple

Turkish	ITUMorphTrans4ET	Google	Bing	Tureng	Yandex
gidebilirse	**if he can go** ✓	can go	if he can go ✓	go	he could leave
gidecek	**he will go** ✓	will go	will go	be destined for	go
gitmelilerdi	**they should have gone** ✓	they should go	they need to get it	go	they have to go
gitmişlerse	**if they had gone**✓	they have gone	if they went	go	if they're gone
gidebilecekse	**if he will be able to go**✓	go abilecekse	can go	go	if you can go
yapmalılarsa	**if they should do** ✓	sign mAlIlArsA	if they're making	go	do they
masalarımızla	**with our tables** ✓	our table	our table	table	our desks

English	ITUMorphTrans4ET	Google	Bing	Yandex
if he can go	**gidebilirse**✓	O gidebilirsiniz	Eğer gidebilir	eğer gidip o
he will go	**gidecek** ✓	O gidecek ✓	o-ecek gitmek	gidecek ✓
they should have gone	**gitmelilerdi**✓ **gitmelilermiş**✓	Onlar gitmiş olmalı	Onlar gitmiş olmalı	gitmelilerdi ✓
if they had gone	**gitmişlerse**✓ **gitmişseler**✓	onlar gitmişti eğer	Onlar ne gitseydin	eğer gitmiş olsalardı
if he will be able to go	**gidebilecekse**✓	O gitmek mümkün olacak eğer	Eğer o-ecek var olmak güçlü-e doğru gitmek için	eğer gitmek mümkün olacak
if they should do	**yapmalılarsa**✓ **etmelilerse**✓	Onlar yapmalıyım	onlar yoksa	eğer yapmalıyım eğer
with our tables	**listelerimizle**✓ **tablolarımızla**✓	Bizim tablolarla ✓	Bizim tablolarla ✓	bizim tablolar ile ✓

Table 2: Comparison of ITUMorphTrans4ET with other popular systems

words. This doesn't mean that the input may be any English utterance but rather English phrases or sentences which maps to single words in the Turkish side or some compound verb forms such as "telefon etmek" (*to phone*). This FST also contains the list of irregular words for correct transformations: e.g. "to go" with lexical form "go+V2+I" will be converted to "went+I" as the surface form whereas a regular verb "play+V2+I" will be converted to "played+I". There are some additional rules for specific cases such as "clap+ing" which be transferred to "clapping" requiring a character repetition of the letter "p". Finally after obtaining the last surface form such as "playing+am+I", we output this in reverse order[5] by the use of a script.

Table 2 gives the comparison of our proposed system with popular online translation systems of namely Google, Yandex, Bing and Tureng. Since the Tureng MT system is only available from Turkish to English, its results are not provided in the second half of the table. The acceptable translations for each case are marked with check marks in the table. As can be noticed, the proposed system produces better results at word-level translation.

4 Learning Use Cases

Learning a foreign language which belongs to a different language family than the native one as in the case of Turkish and English is a problematic task. Since, there are no stable working translators which may be used as a reference, learning becomes a challenging process for new learners. Most commonly used machine translators get use of statistical methods and do not always produce grammatically correct results. In our study, we focus on obtaining a better learning language system which translates words between Turkish and English in two-way effi-

[5]The order of the words are rearranged so that the resulting sentence is grammatical.

Figure 2: An example output from ITUMorphTrans4ET (from English to Turkish)

ciently by providing a user friendly interface for learners.

4.1 Turkish Learning

Learning Turkish is very tough process especially for learners whose mother tongue is not agglutinative due to the fixed order of suffix concatenations to the end of the words which may become dizzy for the new language learners. To give an example for this problem, instead of saying "gitmişlerdi" (*they had been gone*) one may say "gittilermiş" which is not a valid word in terms of suffix order. In the future, ITUMorphTrans4ET system may detect these mistakes (by the help of additional spelling suggesters) and make produce correct translations. In other term, after having a valid analysis and necessary lexical rules, our system can translate any written Turkish word to English language or vice-versa. We believe that same approach may be used for other agglutinative languages by constructing necessary transfer rules.

Figure 2 shows the preliminary interface of ITUMorphTrans4ET where one may type an English lemma and then select morphological properties by using list boxes below. For this example, the word "go" is typed and its properties are selected as "Verb" for the word type, "Positive" for the polarity and so on. Using these information, our system easily translates the word to "gitmeliydim" (*I should have gone*). We believe by improving our system with a more user-friendly interface, the effect of suffixes in Turkish may be efficiently realized and learned by the users.

4.2 English Learning

As explained in previous sections, learning English is as hard as learning Turkish for native Turkish speakers and has same challenging problems. There exists no translation system that works well from Turkish to English at word level. As a consequence, these translators can not be efficient for language learning purposes. However, ITUMorphTrans4ET presents a very strong translation mechanism for two way Turkish-English word translation. It uses morphological model of words in order to translate words. Using the advantage of the morphological structure, even very complex words can be simplified into meaningful tags and then translated to English. Figure 3 gives an example screen for English learners: The word "gitmeliydim" is translated into "I should have gone".

5 Conclusion & Future Work

In this study, we presented our elementary system to develop an educational application for foreign language learners from Turkish and English language pair. Our system uses finite state transducer technology to help the language learning to learn the morphologically complex structure of an agglutinative language and may be applied to other similar languages by developing a transfer model. Although we couldn't test with on large data sets due to the unavailability of APIs of the used MT systems, our preliminary experiments revealed the better performance of ITUMorphTrans4ET. ITUMorphTrans4ET

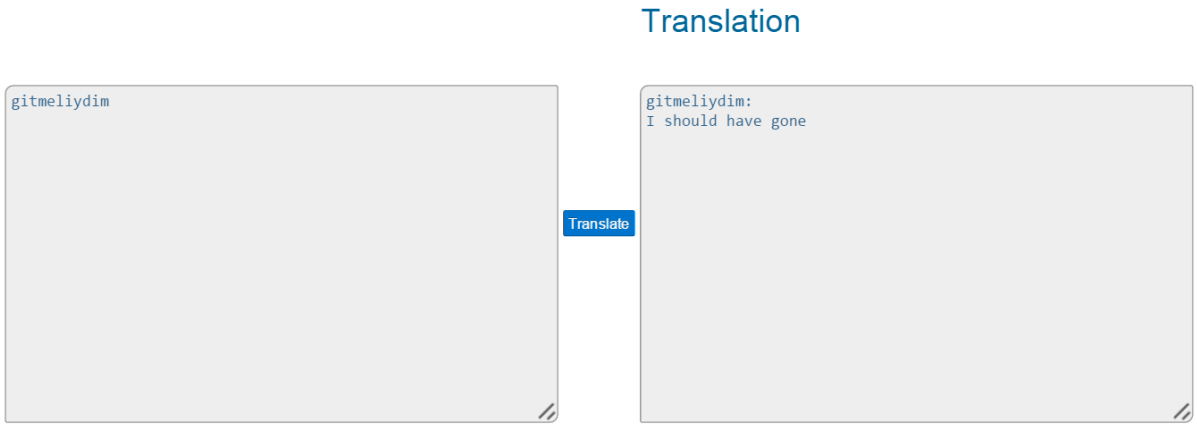

Figure 3: An example output from ITUMorphTrans4ET (from Turkish to English)

may be used in many platforms such as smart boards in classrooms, mobile applications etc. To this end, for future work we plan to focus on improving our system by 1) Developing an user-friendly interactive application for foreign Turkish learners, 2) Developing a mobile application for accessing the interface more easily, and 3) Stepping up to sentence level instead of word level translation by using statistical machine learning approaches. The implementation of ITUMorphTrans4ET is available through a web service found at http://tools.nlp.itu.edu.tr/ (Eryiğit, 2014).

Acknowledgments

The authors want to thank Dilara Torunoğlu Selamet for her valuable contributions to this work.

References

Cyril Allauzen, Michael Riley, Johan Schalkwyk, Wojciech Skut, and Mehryar Mohri. 2007. Openfst: A general and efficient weighted finite-state transducer library. In *Implementation and Application of Automata*, pages 11–23. Springer.

Kenneth R Beesley and Lauri Karttunen. 2003. Finite-state morphology: Xerox tools and techniques. *CSLI, Stanford*.

Arianna Bisazza and Marcello Federico. 2009. Morphological pre-processing for Turkish to English statistical machine translation. In *IWSLT*, pages 129–135.

Stephen Pit Corder. 1983. A role for the mother tongue. *Language transfer in language learning*, pages 85–97.

Aydin Y. Durgunoglu and Barbara J. Hancin-Bhatt. 1992. The role of first language in the second-language reading process. Technical report, University Illinois at Urbana-Champaign.

Ilknur Durgar El-Kahlout and Kemal Oflazer. 2010. Exploiting morphology and local word reordering in English-to-Turkish phrase-based statistical machine translation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(6):1313–1322.

Gülşen Eryiğit. 2014. ITU Turkish NLP web service. In *Proceedings of the Demonstrations at the 14th Conference of the EACL*, Gothenburg, Sweden, April. Association for Computational Linguistics.

Elif Eyigöz, Daniel Gildea, and Kemal Oflazer. 2013. Simultaneous word-morpheme alignment for statistical machine translation. In *Proceedings of NAACL-HLT*, pages 32–40.

Kenji Hakuta, Yuko Goto Butler, and Daria Witt. 2000. How long does it take English learners to attain proficiency? *University of California Linguistic Minority Research Institute*.

Kenji Hakuta. 1999. A critical period for second language acquisition? a status review. *National Center for Early Development of Learning. Chapel Hill, NC: University of North Carolina*.

Krister Lindén, Miikka Silfverberg, and Tommi Pirinen. 2009. Hfst tools for morphology–an efficient open-source package for construction of morphological analyzers. In *State of the Art in Computational Morphology*, pages 28–47. Springer.

Mehryar Mohri. 1997. Finite-state transducers in language and speech processing. *Computational linguistics*, 23(2):269–311.

Håkan Ringbom. 1987. *The role of the first language in foreign language learning*, volume 34. Multilingual Matters Ltd.

Emmanuel Roche and Yves Schabes. 1997. *Finite-state language processing*. MIT press.

Muhammet Şahin, Umut Sulubacak, and Gülşen Eryiğit. 2013. Redefinition of Turkish morphology using flag diacritics. In *Proceedings of The Tenth Symposium on Natural Language Processing (SNLP-2013)*, Phuket, Thailand, October.

Michael Swan. 1997. The influence of the mother tongue on second language vocabulary acquisition and use. *Vocabulary: Description, acquisition and pedagogy*, pages 156–180.

Reyyan Yeniterzi and Kemal Oflazer. 2010. Syntax-to-morphology mapping in factored phrase-based statistical machine translation from English to Turkish. In *Proceedings of the 48th Annual Meeting of ACL*, pages 454–464. Association for Computational Linguistics.