Machine Translation of languages and dialects

Parneet Kaur[1], Simrat Kaur[2]

[1]Student, Computer Science and Engineering,Baba Banda Singh Bahadur Engineering College, Punjab, India
[2]Assistant Professor, Dept. of Computer Science and Engineering,Baba Banda Singh Bahadur Engineering College, Punjab, India

---***---

Abstract - *Machine translation processes a natural language to translate it to another natural language. Machine Translation activities in India are relatively young and their demand is increasing due to increased exchange of information on internet across the world, due to which machine translation has become an important research subfield under Artificial Intelligence. India is very rich in linguistics. It has 22 official languages and further 2000 dialects of the regional languages written in 12 different scripts, whose speakers are greater than 1.25 billion. Thus there is a great need of translation among these languages to make someone understand the information written in an another local language not known to the former. This paper categorizes the machine translation systems for Indian languages on basis of the approach used and source language used and then focus on dialectal machine translation.*

Key Words: Machine Translation, Example-based MT, Interlingua-based MT, Transfer-based MT, Dialects.

1.INTRODUCTION

India is a multilingual country where language changes every 50 miles. There are 22 official languages in India and approximately 2000 dialects spoken by different communities [1]. In such a country where so many languages are used as official languages, machine translation has become a necessity as manual translation from one official language to another is very costly and time consuming. State governments do their documentation work in their own regional language while centre government use English or Hindi for documentation and the many newspapers are also printed in regional languages. So the need of translation has become more. Also internet is used as source of information where anyone can share the information and anyone can be benefitted from that information. Recently the rate of written colloquial text has increased dramatically. It is being used as a medium of expressing ideas especially across the WWW, usually in the form of blogs and partially colloquial articles [2]. Hence to get benefitted, the colloquial text must be understood to all, i.e., it has to be translated to reader understandable language or to a language which all can understand. Hence there is a great need of good machine translation systems that can translate between regional languages and further between dialects of a regional language.

A good deal of work has been done for translation Hindi to English, English to Hindi and other regional languages to English or Hindi but there are very few machine translation systems in India for translation among the dialects of a language. Institutions like IIT Kanpur, IIT Bombay, IIIT Hyderabad, University of Hyderabad, NCST Mumbai, CDAC Pune, CDAC Noida, Department of Computer Science and Engineering Jadavpur University, Kolkata, JNU New Delhi etc are playing a major role in developing the MT systems in India [1]. This paper summarizes the machine translation systems made for Indian languages and machine translation systems for dialects across the world.

This paper is divided into four sections. Second section discusses machine translation systems for dialects of different languages across the world. Third section summarizes the second section in a table. Fourth section is the conclusion.

2. MACHINE TRANSLATION SYSTEMS FOR LANGUAGE DIALECTS

3.1 Cantonese-Mandarin Dialect MT System (2002)

The system is designed for conversion two dialects of Chinese language. The system takes Cantonese sentence as input and give output in Mandarin sentence. The system have linguistic rules for syntax conversion, a collocation list for disambiguation, and a bilingual dictionary of Cantonese and Mandarin words for substitution. As there is much less serious difference in between two dialects in comparison to two languages, machine translation is considered more practical here. Dialect MT contains a bilingual dictionary of about 10,000 words and some rules and a corpus of Cantonese Mandarin dialect [3].

3.2 Magead (2006)

MAGEAD is a morphological analyzer and generator for the Arabic language family. The system addresses both the Modern Standard Arabic and the spoken dialects i.e. it processes the morphology of Arabic dialects also. The system relates a lexeme and a set of linguistic features to a surface word form through a sequence of transformations. It has 69 Morphophonemic/phonological rules and 53 orthographic rules to rewrite the word. The system is the extension of

Kiraz (2000) which has four tiers and fifth tier is added to it. The system used Levantine as its first dialect. The system can be used for new dialects in absence of a lexicon and with a restrained amount of manual knowledge engineering needed [4].

3.3 A Hybrid Approach For Converting Written Egyptian Colloquial Dialect into Diacritized Arabic (2008)

The system converts a written Egyptian colloquial sentence to a diacritized Modern Standard Arabic sentence. Egyptian Colloquial dialect has been choosen because of its more use in blogs and articles over the internet. The resources are collected using a rule based approach from a large amount of data across the WWW. The system contains a lexicon of 41705 words out of which 9085 are non MSA words, 3000 distinct colloquial words and rest are spelling mistake words and non Arabic names. The system is trained to distinguish between MSA words and Colloquial words. POS tagging is done using Statistical approach and then Rule based approach is used to convert Egyptian Arabic words to their corresponding MSA words. The system is tested for 1000 words and accuracy for converting colloquial to MSA word is 88% [2].

3.4 Arabic Dialect Handling in Hybrid Machine Translation (2010)

The system is an extension of a Hybrid Machine Translation System for handling Arabic dialects. It uses a Statistical decoder which contains four types of rules-lexical, syntactic, argument structure, and functional structure rules, semantic disambiguation information, a statistical bilingual lexicon, bilingual phrase table and target language models. The system is tested with and without dialect normalization against BLEU score and result is higher score with dialect normalization [5].

3.5 Enhancement of Morphologiacal Analyzer With Compound, Numeral and Colloquial Word Handler (2011)

This system translates written colloquial Tamil into written normalise formal Tamil. A Rule based approach is used for handling compounds and numerals and a pattern mapping based approach is used for handling colloquial words [6].

3.6 The Arabic Online Commentary Dataset (2011)

This system holds 52M –word monolingual dataset which is rich in dialectal content. Also the system is trained to identify the dialectal content and to specify the level of dialectal content in a sentence. The data is extracted from three newspapers which contained high degree of dialectal content from Levantine, gulf and Egyptian dialects. The system can distinguish the dialectal content from MSA and from other dialectal content [7].

3.7 Dialectal to Standard Arabic Paraphrasing to improve Arabic-English Statistical Machine Translation (2011)

This project was supported by DARPA GALE program. An existing MSA analyzer is extended by adding dialectal out of vocabulary (OOV) words and low frequency words. The system produces Standardized paraphrases in MSA. Two dialect varieties has been used-Levantine and Egyptian. A light rule based approach is used. To generate the paraphrase lattices 11 morphological rules were used. This system improves the BLEU score on blind test by 0.56 absolute BLEU. It gives correct translation in 74% of the time for OOVs and 60% of the time for low frequency words [8].

3.8 Unidic (2012)

This work was partially supported by the collaborative research project "Study of the history of the Japanese language using statistics and machine-learning" carried out at the National Institute for Japanese Language and Linguistics. It is an electronic dictionary for Early Middle Japanese or classical Japanese. The accuracy of the system for analysing Japanese Classical text is 97% [9].

3.9 Machine Translation of Arabic Dialects (2012)

The work was supported partially by DARPA/IPTO and partially by EuroMatrixPlus project funded by European Commission. The system translates Levantine and Egyptian dialects to English. Data was collected by using crowdsourcing technique consisting 1.1M words of Levantine and 380K of Egyptian dialect. The system is trained on 1.5M of dialectal data performs 6.3 to 7.0 BLEU points higher than a Modern Standard Arabic MT system trained on 150M-word Arabic-English parallel corpus [10].

3.10 Sentence Level Dialect Identification for Machine Translation System Selection (2012)

This system improves the output of different previously developed MT systems by selecting what sentence go to which MT system. The system consider two dialects-Levantine and Egyptian along with MSA. This system can identify the type of sentence if it is a MSA only sentence or include some dialectal content so that corresponding suitable MT system can be used and the accuracy for the output increases. This best system selection approach improves over the best baseline single MT system by 1.0% absolute BLEU point on a blind test set [11].

3.11 Automatic Conversion of Dialectal Tamil to Standard Written Tamil Text Using FSTs (2014)

The system can translate various spoken Tamil dialects to Standard Written Tamil text. Finite State Transducers are used for obtaining equivalent Standard Tamil words and Conditional Random Fields are used for handling agglutination and compounding in the resultant text. The system can translate central Tamil, Madurai Tamil, Tirunelveli Tamil, Brahmin tamil, kongu Tamil and common spoken forms. The translation accuracy is higher for Kongu Tamil dialect and lower for Madurai and Tirunelveli due to polysemous nature of the words of these dialects [12].

3.12 Domain and Dialect Adaptation for Machine Translation into Egyptian Arabic (2014)

The system translates English to Egyptian Dialect of Arabic language by first translating English to MSA and the MSA to Egyptian Arabic. Translation from English to MSA is done using a large bilingual corpus and translation from MSA to Egyptian is done using two pathways- two step domain and dialect adaptation and one step simultaneous domain and dialect adaptation. The system uses 100k sentence tri-parallel corpus of English, MSA, and Egyptian Arabic generated by a rule-based transformation. The system translates better with two step domain with a BLEU score of 42.9 [13].

3.13 Handling OOV Words in Dialectal Arabic to English Machine Translation (2014)

This work was supported by the Defense Advanced Research Projects Agency (DARPA), the BOLT program with subcontract from Raytheon BBN. The system replaces the OOV(Out of Vocabulary) dialectal words with Standard Written MSA to enhance the Statistical Machine translation of Arabic to English. Two dialect identification MT Systems-AIDA and MADAMIRA were used to identify and replace OOV words and the output is fed to Statistical Arabic English MT System. This system enhances BLEU score of Arabic English MT System by 0.4% using AIDA and 0.3% using MADAMIRA [14].

3.14 Dialect Resolution: A Hybrid Approach (2014)

The system translates informal sentences and slangs of Thrissur dialect to a formal format. A hybrid approach mixing Rule bas and machine learning approaches is used. Accuracy in the following target words are depends upon the previous resolved formal words [15].

3.15 Punjabi Dialects Conversion System For Malwai and Doabi Dialects (2015)

The system translates sentences between two dialects of Punjabi language-Malwai and Doabi, and from Standard Punjabi to these also. A Rule base approach is used with three bilingual dictionaries that translates Standard Punjabi to Malwai, Standard Punjabi to Doabi, Malwai to Doabi and Doabi to Malwai. Accuracy of the system for Standard Punjabi to Malwai is 95% and Standard Punjabi to Doabi is 94% [16].

4.SUMMARY

The following table summarizes the above Dialectal MT Systems for their features.

Sr. No.	System Name	Language	Year
1	CANTONESE-MANDARIN DIALECT MT SYSTEM	Cantonese to Mandarin	2002
2	MAGEAD	Levantine Arabic and MSA	2006
3	A HYBRID APPROACH FOR CONVERTING WRITTEN EGYPTIAN COLLOQUIAL DIALECT INTO DIACRITIZED ARABIC	Egyptian to Modern Standard Arabic	2008
4	ARABIC DIALECT HANDLING IN HYBRID MACHINE TRANSLATION	15 Colloquial Arabic Dialects to MSA	2010
5	ENHANCEMENT OF MORPHOLOGICAL ANALYZER WITH COMPOUND, NUMERAL AND COLLOQUIAL WORD	Colloquial Tamil to formal Tamil	2011

	HANDLER		
6	THE ARABIC ONLINE COMMENTARY DATASET	Levantine, Gulf and Egyptian	2011
7	DIALECTAL TO STANDARD ARABIC PARAPHRASING TO IMPROVE ARABIC-ENGLISH STATISTICAL MACHINE TRANSLATION	Levantine and Egyptian to MSA	2011
8	UNIDIC	Early Middle Japanese	2012
9	MACHINE TRANSLATION OF ARABIC DIALECTS	Levantine and Egyptian to English	2012
10	SENTENCE LEVEL DIALECT IDENTIFICATION FOR MACHINE TRANSLATION SYSTEM SELECTION	Levantine, Egyptian and MSA	2012
11	AUTOMATIC CONVERSION OF DIALECTAL TAMIL TEXT TO STANDARD WRITTEN TAMIL TEXT USING FSTs	central Tamil, Madurai Tamil, Tirunelveli Tamil, Brahmin tamil, kongu Tamil and common spoken forms to Standard Tamil	2014
12	DOMAIN AND DIALECT ADAPTATION FOR MACHINE TRANSLATION INTO EGYPTIAN ARABIC	English to Egyptian	2014
13	HANDLING OOV WORDS IN DIALECTAL ARABIC TO ENGLISH MACHINE TRANSLATION	Dialectal Arabic to English	2014
14	DIALECT RESOLUTION: A HYBRID APPROACH	Informal Thrissur dialect to formal Thrissur	2014
15	PUNJABI DIALECTS CONVERSION SYSTEM FOR MALWAI AND DOABI DIALECTS	Standard Punjabi to Malwai and Doabi, Malwai to Doabi, Doabi to Malwai	2015

4. CONCLUSIONS

This paper describes MT Systems for Indian languages in brief and MT Systems for Dialectal processing in longitudinal and latitudinal way. There are good translation Systems for Indian languages which translates to English but there is a large room for the dialect processing as there is very less work is done on Indian Dialects and India has 2000 dialects out of which only a few number of dialects are considered for machine Translation. Dialect processing will ease the information retrieval from the internet.

REFERENCES

[1] G. V. Garje and G. K. Kharate, “Survey of machine Translation Systems in India”, International Journal

on Natural Language Computing (IJNLC) Vol. 2, No.4, October2013, pp 47-67.

[2] Hitham M. Abo Bakr, et al., "A hybrid approach for converting written Egyptian colloquial dialect into diacritized Arabic", 6th International Conference on Informatics and Systems, INFOS2008, Cairo, Egypt,2008.

[3] Zheng, X. and Kowloon, H. H., "Dialect MT- A Case Study between Cantonese and mandarin", Proceedings of 17th International Conference on Computional Linguistics, May 2002, DOI: 10.3115/980691.980807.

[4] Nizar, H. and Rambow, O., "A Morphological Analyzer and Generator fot the Arabic Dialects", Proceedings of the 21st International Conference on Computional Linguistics and 44th Annual Meeting of the ACL, Sydney, July 2006, pp 681-688.

[5] Sawaf, H., "Arabic Dialect Handling in Hybrid machine Translation", In Proceedings of the Conference of the Association of machine Translation in the Americas(AMTA), Denver, Colorado.

[6] Ranganathan, K., et al., "Enhancement of Morphological Analyzer with Compound, Numeral and Colloquial Word Handler", Proceedings of ICON-2011: 9th International Conference on Natural Language Processing, 2011.

[7] Zaidan, F. O., Burch, C. C., "The Arabic Online Commentary Dataset: an Annotated Dataset of Informal Arabic with High Dialectal Content", Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: shortpapers, Portland, Oregon, June 19-24, 2011, pp 37–41.

[8] Salloum, W. and Nizar Habash, N., "Dialectal to Standard Arabic Paraphrasing to Improve Arabic-English Statistical Machine Translation", Proceedings of EMNLP 2011, Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, UK, July 27–31, 2011, pp 10–21.

[9] Ogiso, T., et al., "UniDic for Early Middle Japanese: a Dictionary for Morphological Analysis of Classical Japanese",In LREC 2012 Proceedings, may 2012.

[10] Zbib, R., et al., "Machine Translation of Arabic Dialects", 2012 Conference of the North American Chapter of the Association for Computational linguistics: Human Language technologies, Montreal, Canada, June 3-8, 2012, pp 49-59.

[11] Salloum, W., et al., "Sentence level Dialect Identification for machine translation System Selection", Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), Baltimore, Maryland, USA, June 23-25 2014, pages 772–778.

[12] Marimuthu, K. and Devi, S. L., " Automatic Conversion of Dialectal Tamil Text to Standard Written Tamil text using FSTs", Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM, Baltimore, Maryland USA, June 27, 2014, pp 37-45.

[13] Jeblee, S., et al., "Domain and Dialect Adaptation for Machine Translation into Egyptian Arabic", Proceedings of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP), Doha, Qatar, October 25, 2014, pp 196–206.

[14] Minian, A. M., et al., "Handling OOV Words in Dialectal Arabic to English Machine Translation", Language Technology for Closely Related Languages and Language Variants (LT4CloseLang), Doha, Qatar, October 29, 2014, pp 99–108.

[15] Sarath, K. S., et al., "Dialect resolution: A Hybrid Approach", An International Journal of Engineering Sciences, Special Issue iDravadian, December 2014, Vol. 15 ISSN: 2229-6913.

[16] Singh, A. and Singh, P., "Punjabi Dialects Conversion System for Malwai and Doabi Dialects", International Journal of Science and Technology, October 2015, Vol. 8(27), ISSN : 0974-6846.