PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA

MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

UNIVERSITY MUSTAPHA STAMBOULI OF MASCARA

Faculty of Exact Sciences

Department of Computer Science

Dissertation

Submitted in partial fulfilment of the requirements for Doctorate degree in Computer Science

Option: Artificial Intelligence

Theme

Neural Machine Translation for Arabic Media Content

Presented by **Baligh BABAALI**

Jury:

President	Aaaaa AAAAA	Professor	University of Mascara
Director	Mohammed SALEM	Professor	University of Mascara
Examiner	Ddddd DDDDD	Professor	University of Mascara
Examiner	Eeeee EEEEE	Associate Prof.	University of Mascara
Examiner	Fffff FFFFF	Associate Prof.	University of Mascara

Month 2024

Acknowledgements

Dedication

ملخص

ملخص .

كلمات مفتاحية: ملخص، ملخص، ملخص، ملخص، ملخص.

Abstract

Abstract Abstract.

Key words: *Abstract, Abstract, Abstract, Abstract, Abstract.*

Contents

List of Figures

List of Tables

List of Abbreviations

BT: Back Translation
FFN: Feed-Forward Network
GRU: Gated Recurrent Units
LLM: Large Language Model
LSTM: Long Short-Term Memory
MT: Machine Translation
NLP: Natural Language Processing
NMT: Neural Machine Translation
RNN: Recurrent Neural Network
SMT: Statistical Machine Translation

Chapter 1

Introduction

Here goes the Introduction.

1.1 Problem Identification and Motivation

1.2 Objectives and Scope

1.3 Contributions

1.4 Outline of the Thesis

PART I:
BACKGROUND AND RELATED WORK

Chapter 2

Background

2.1 Overview

2.2 Arabic Language

2.3 Social Media

2.4 Natural Language Processing

2.5 Machine Translation

Machine Translation is a procedure that uses computer pieces of software to express text from one natural language NL (SL i.e. source language) in another NL (TL i.e. target language). In any human or automated translation process, the meaning of the source sentences must be fully reproduced into the target translated sentences, which is only simple on the surface.

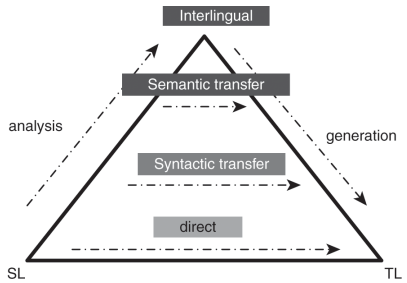

Figure 2.1: The Vauquois triangle, illustrating the foundations of machine translation.

The different approaches to MT fall into three categories: methods that depend on rules and knowledge (linguistic-based). Approaches that are empirical and data-driven (corpus-

based); and finally, hybrid methods.

2.5.1 Linguistic Approaches

These MT approaches attempt to formalize all the necessary knowledge required for translation, using expert methods. The "Vauquois triangle" presented in Figure 2.1 is a generic representation of these techniques.

2.5.1.1 Direct Approach

or Direct MT (DMT) is, the simplest MT approach. It operates at the word level, i.e. the words' translation is done word by word, just, as a dictionary does, and generally without much correspondence of their meaning [1].

2.5.1.2 Rule-based MT

(RBMT) uses linguistic knowledge of source and target languages fundamentally collected from (bilingual) dictionaries and grammars encompassing the principal morphological, syntactic and/or semantic rules of each language respectively [1]. RBMT approach suffers from the impossibility of writing all the rules of all the languages, because this task requires large and important linguistic knowledge.

2.5.1.3 Interlingual MT

The term "Interlingua" refers to a language that serves as a bridge between two languages. In this method, SL is turned into an assistant/mediator language (representation) which is independent of the languages concerned by the translation. This auxiliary form is then used to specify the TL's translated verse. This approach focuses on a single representation for different languages [2].

2.5.1.4 Transfer-based MT

(TBMT) is similar to Interlingual-MT in that it generates a translation from an intermediate structure that mimics the original sentence's meaning. The source text is translated into a less language-specific intermediate representation. This form is then translated into a target language structure with a comparable structure, and the text is generated in the target language. The source and target languages' morphological, syntactic, and/or semantic information is used in the transfer process. As a result, TBMT can make use of knowledge of both the source and target languages. [3].

2.5.2 Corpus Approaches

Corpus techniques use empirical methods to ensure that all linguistic knowledge is learned empirically and automatically from corpora, which are collections of parallel datasets of source and target phrases that are translated to each other.

2.5.2.1 Example-based MT

The main idea behind (EBMT) is analogy [4]. The primary concept is to build new translations on top of current examples. Bilingual parallel corpora containing sentence pairs are used to train EBMT systems. It's used to translate similar-sounding sentences by looking for the closest source example to the source word or phrase in parallel corpora. Nagao has appropriately classified this procedure into three steps [4]:

- Fragments are matched against a database of real examples.
- Identifying the translation fragments that correlate (Alignment)
- Putting these together to create the target text

2.5.2.2 Statistical MT

(SMT) generates translation hypotheses in a target language t based on a sentence in a source language s with the highest conditional probability $P(t|s)$ [5, 6]. The translation direction will be inverted to a translation model (TM) $P(s|t)$ and a language model (LM) $P(t)$ will be included by applying the Bayes rule. The following equation (2.1) is used to optimize the likelihood of the best translation:

$$t_{best} = arg_t max(P(t|s)) = arg_t max(P(s|t) \times P(t)) \tag{2.1}$$

where $P(s|t)$ is the TM and $P(t)$ is the LM.

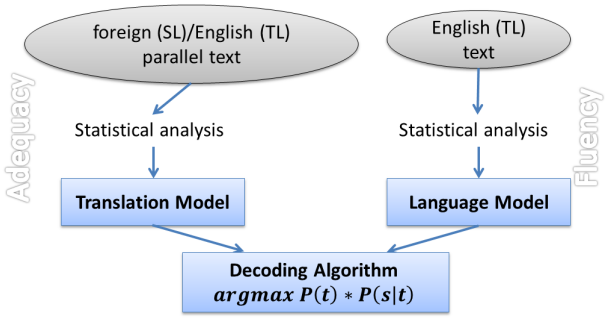

Figure 2.2: Statistical Machine Translation approach

SMT requires a language model, a translation model, and a decoding method in general. The TM, on the one hand, assures that the target hypothesis created matches the source sentence. The LM, on the other hand, ensures that the output is grammatically correct (Figure 2.2).

2.6 Neural Networks

2.6.1 Feed-Forward Neural Networks

2.6.2 Recurrent Neural Networks

2.6.2.1 Stacked Recurrent Neural Networks

2.6.2.2 Bidirectional Recurrent Neural Networks

2.6.3 Long Short-Term Memory and Gated Recurrent Units

2.6.4 Transformers

2.6.5 Large Language Models

2.7 Neural Machine Translation

2.7.1 Neural Machine Translation with Recurrent Neural Networks

2.7.2 Neural Machine Translation with Transformers

2.7.3 Training Neural Machine Translation Models

2.7.4 Decoding

2.8 Evaluation

2.8.1 Human Evaluation

2.8.2 Automatic Evaluation

2.9 Summary

Chapter 3

Related Work

3.1 Overview

3.2 Summary

PART II:
DATASET CREATION

Chapter 4

Bilingual and Monolingual Corpora

4.1 Overview

4.2 Bilingual Corpora

4.2.1 Existing Parallel Corpora

4.2.2 Data Sources

4.2.3 Preprocessing

4.3 Monollingual Corpora

4.3.1 Existing Monolingual Corpora

4.3.2 Data Sources

4.3.3 Preprocessing

4.4 Data Augmentation

4.4.1 Back Translation Augmentation

4.4.2 Right Rotation Augmentation

4.4.3 Named Entities Replacement Augmentation

4.5 Summary

Chapter 5

Data Augmentation

5.1 Overview

5.2 Data Augmentation

5.2.1 Back Translation Augmentation

5.2.2 Right Rotation Augmentation

5.2.3 Named Entities Replacement Augmentation

5.3 Summary

PART III:
MODELS CONSTRUCTION AND EVALUATION

Chapter 6

Seq2Seq Neural Machine Translation

6.1 Overview

6.2 System Architecture

6.3 Baseline System

6.4 Experiments and Evaluation

6.5 Results and Discussions

6.6 Summary

Chapter 7

LLM-based Machine Translation

7.1 Overview

7.2 System Architecture

7.3 Baseline System

7.4 Experiments and Evaluation

7.5 Results and Discussions

7.6 Summary

Chapter 8

Conclusion

8.1 Summary

Here goes the conclusion.

8.2 Limitations

8.3 Future work

Bibliography

[1] M. D. Okpor. Machine translation approaches: Issues and challenges. *IJCSI International Journal of Computer Science Issues*, 11(2):159–165, Sep 2014.

[2] Neeha Ashraf and Manzoor Ahmad. Machine translation techniques and their comparative study. *International Journal of Computer Applications*, 125(7):25–31, Sep 2015. Published by Foundation of Computer Science (FCS), NY, USA.

[3] Thi-Ngoc-Diep DO. *Extraction de corpus parallèle pour la traduction automatique depuis et vers une langue peu dotée*. PhD thesis, UNIVERSITÉ DE GRENOBLE, 2011.

[4] Makoto Nagao. Framework of a mechanical translation between japanese and english by analogy principle. *Artificial and Human Intelligence, A. Elithorn and R. Banerji (eds.) North-Holland*, pages 173–180, 1984.

[5] Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederick Jelinek, Robert L Mercer, and Paul Roossin. A statistical approach to language translation. In *Coling Budapest 1988 Volume 1: International Conference on Computational Linguistics*, 1988.

[6] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[7] Wikipedia. *Machine translation*, 2020 (accessed February 3, 2020).

[8] Sadik Bessou. *Contribution au Niveau de l'Approche Indirecte à Base de Transfert dans la Traduction Automatique*. PhD thesis, University FERHAT ABBAS, Setif 1, Jun 2015.

[9] Francisco Guzmán, Houda Bouamor, Ramy Baly, and Nizar Habash. Machine translation evaluation for Arabic using morphologically-enriched embeddings. In *Proceedings*

of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1398–1408, Osaka, Japan, dec 2016. The COLING 2016 Organizing Committee.

[10] Baijun Ji, Zhirui Zhang, Xiangyu Duan, Min Zhang, Boxing Chen, and Weihua Luo. Cross-lingual pre-training based transfer for zero-shot neural machine translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:115–122, 04 2020.

[11] Abdelhadi Soudi, Violetta Cavalli-Sforza, and Abderrahim Jamari. Prototype english-to-arabic interlingua-based mt system. In *Proceedings of the Third International Conference on Language Resources and Evaluation: Workshop on Arabic Language Resources and Evaluation: Status and Prospects*, pages 18–25, Las Palmas de Gran Canaria, Spain, 01 2002.

[12] Dimitra Anastasiou. *Idiom Treatment Experiments in Machine Translation*. PhD thesis, Universität des Saarlandes, Jan 2010.

[13] Haithem Afli. *La traduction automatique statistique dans un contexte multimodal*. PhD thesis, UNIVERSITÉ DU MAINE, Jul 2014.

[14] W. John Hutchins. The georgetown-ibm experiment demonstrated in january 1954. In *Machine Translation: From Real Users to Research*, pages 102–114. Springer, 2004.

[15] Raphaël Rubino. *Traduction automatique statistique et adaptation à un domaine spécialisé*. PhD thesis, UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE, Nov 2011.

[16] Wael Salloum and Nizar Habash. Elissa: A dialectal to standard arabic machine translation system. In *COLING 2012: Demonstration Papers*, number December 2012, pages 385–392, Mumbai, 2012.

[17] Mohamed Ali Sghaier and Mounir Zrigui. Rule-based machine translation from tunisian dialect to modern standard arabic. *Procedia Computer Science*, 176:310–319, 2020.

[18] John Pierce, John Carroll, Eric Hamp, David Hays, Charles Hockett, Anthony Oettinger, and Alan Perlis. Language and machines: Computers in translation and linguistics. Report 1416, National Academy of Sciences/National Research Council, Washington, D. C., 1966.

[19] Thomas Schneider. *Progress in Machine Translation*, chapter 10, pages 99–103. IOS Press, Amsterdam, Netherlands, 1993.

[20] Anne-Marie Loffler-Laurian. *La traduction automatique*, chapter 1, pages 22–23. Presses Universitaires du Septentrion, Villeneuve d'Ascq, France, 1996.

[21] W. John Hutchins. *An Encyclopaedia of Translation: Chinese-English, English-Chinese*, chapter Machine Translation, page 599. The Chinese University Press, 2001.

[22] Ch. Boitet and Patrick Guillaume. Ariane-78: an integrated environment for automated translation and human revision. In *COLING 1982*, 1982.

[23] H.D. Maas. The saarbrüken automatic translation system (susy). In *European Congress on Information Systems and Networks, Overcoming the language barrier*, volume 1, pages 585–592. München, May 1977.

[24] Makoto Nagao, Jun ichi Tsujii, and Jun ichi Nakamura. The japanese government project for machine translation. *Computational Linguistics*, 11(2-3):91–110, Apr-Sep 1985.

[25] Margaret King. Eurotra – a european system for machine translation. *Lebende Sprachen*, 26(1):12–14, 1981.

[26] Sneha Tripathi and Juran Krishna Sarkhel. Approaches to machine translation. *Annals of Library and Information Studies*, 57:388–393, Dec 2010.

[27] Uwe Muegge. An excellent application for crummy machine translation: Automatic translation of a large database. In *Annual Conference of the German Society of Technical Communicators*, pages 18–21, 2006.

[28] Kareem Darwish. Arabizi detection and conversion to arabic. In *Proceedings ofthe EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP)*, pages 217–224, Doha, Qatar, 2015.

[29] Bonnie J. Dorr. Unitran: An interlingual approach to machine translation. In *AAAI-87 Proceedings*, pages 534–539, 1987.

[30] Chiew Kin Quah. *Machine Translation Systems*, chapter Machine Translation Systems, pages 57–92. Palgrave Macmillan UK, London, 2006.

[31] Bonnie J. Dorr. Interlingual machine translation a parameterized approach. *Artificial Intelligence*, 63(1):429–492, 1993.

[32] Yehoshua Bar-Hillel. The Present Status of Automatic Translation of Languages. *Advances in Computers*, 1(C):91–163, 1960.

[33] Harold Somers. Review article: Example-based machine translation. *Machine Translation*, 14(2):113–157, 1999.

[34] Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2020.

[35] Andreas Eisele and Yu Chen. MultiUN: A multilingual corpus from united nation documents. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2868–2872, Valletta, Malta, 2010. European Language Resources Association (ELRA).

[36] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT Summit X*, pages 79–86, 2005.

[37] Michel Simard. The baf: A corpus of english-french bitext. In *First International Conference on Language Resources & Evaluation*, volume 1, pages 489–494, Granada, Spain, May 1998.

[38] Satoshi SATO and Makoto NAGAO. Toward memory-based translation. In *Proc. 13th International Conference on Computational Linguistics (Coling 90)*, pages 247–252, Helsinki, Finland, 1990.

[39] Anne Osherson and Christiane Fellbaum. The representation of idioms in wordnet. In *Principles, Construction and Application of Multilingual Wordnets. Proceedings of the Fifth Global WordNet Conference (GWC 2010), Mumbai, India. Narosa Publishing House*, 2010.

[40] Bogdan Babych and A. Hartley. Improving machine translation quality with automatic named entity recognition. 2003.

[41] Neeraj Agrawal and Ankush Singla. Using named entity recognition to improve machine translation. *Technical report, Standford University, Natural Language Processing*, 2012.

[42] Ulf Hermjakob, Kevin Knight, and Hal Dauḿe. Name translation in statistical machine translation learning when to transliterate. *ACL-08: HLT - 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, (06):389–397, 2008.

[43] Diego Moussallem, Matthias Wauer, and Axel Cyrille Ngonga Ngomo. Semantic web for machine translation: Challenges and directions. *CEUR Workshop Proceedings*, 2576:1–9, 2019.

[44] Mirjam Sepesy Maučec and Gregor Donaj. Machine translation and the evaluation of its quality. In *Recent Trends in Computational Intelligence*. IntechOpen, 2019.

[45] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, Jul 2002. Association for Computational Linguistics.

[46] Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. The penn arabic treebank: Building a large-scale annotated arabic corpus. *NEMLAR Conference on Arabic Language Resources and Tools*, 01 2004.

[47] Ibrahim Abu El-khair. 1.5 billion words arabic corpus. *arXiv e-prints*, nov 2016. Provided by the SAO/NASA Astrophysics Data System.

[48] Robert Parker, David Graff, Ke Chen, Junbo Kong, and Kazuaki Maeda. Arabic gigaword fifth edition LDC2011T11.

[49] El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. Investigating code-mixed modern standard arabic-egyptian to english machine translation. *arXiv preprint arXiv:2105.13573*, 2021.

[50] El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. Turjuman: A public toolkit for neural arabic machine translation. *arXiv preprint arXiv:2206.03933v1*, 2022.

[51] Diadeen Ali Hameed, Tahseen Ameen Faisal, Ali Mustafa Alshaykha, Ghanim Thiab Hasan, and Harith Abdullah Ali. Automatic evaluating of russian-arabic machine translation quality using meteor method. *AIP Conference Proceedings*, 2386(1):040036, 2022.

[52] Laith H. Baniata, Sangwoo Kang, and Isaac. K. E. Ampomah. A reverse positional encoding multi-head attention-based neural machine translation model for arabic dialects. *Mathematics*, 10(19), 2022.

[53] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*, Mar 2017.

[54] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR'15)*, 2015.

[55] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In Association for Computational Linguistics, editor, *Proceedings of the 8th Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST'14)*, pages 103–111, 2014.

[56] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS'14)*, pages 3104–3112, 2014.

[57] Akiko Eriguchi, Spencer Rarrick, and Hitokazu Matsushita. Combining translation memory with neural machine translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 123–130, 2019.

[58] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[59] Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.

[60] Dimitar Shterionov, Pat Nagle Laura Casanellas, Riccardo Superbo, and Tony O'Dowd. Empirical evaluation of nmt and pbsmt quality for large-scale translation production. In *Conference Booklet*, page 74, 2017.

[61] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark, Sep 2017. Association for Computational Linguistics.

[62] Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation:*

Shared Task Papers, pages 272–303, Belgium, Brussels, Oct 2018. Association for Computational Linguistics.

[63] Felix Stahlberg. Neural machine translation: A review and survey.

[64] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning. In Doina Precup and Yee Whye Teh, editors, *Proceedings of Machine Learning Research*, volume 70, pages 1243–1252, International Convention Centre, Sydney, Australia, 06-11 Aug 2017. PMLR.

[65] Haoran Xu, Benjamin Van Durme, and Kenton Murray. Bert, mbert, or bibert? a study on contextualized embeddings for neural machine translation. *arXiv preprint arXiv:2109.04588*, 2021.

[66] Amine Abdaoui, Mohamed Berrimi, Mourad Oussalah, and Abdelouahab Moussaoui. Dziribert: a pre-trained language model for the algerian dialect. *arXiv preprint arXiv:2109.12346*, 2021.

[67] Abir Messaoudi, Hatem Haddad, Moez BenHajhmida, Malek Naski, Ahmed Cheikhrouhou, Nourchene Ferchichi, Abir Korched, Faten Ghriss, and Amine Kerkeni. Tunbert: Pretrained contextualized text representation for tunisian dialect. *arXiv preprint arXiv:2111.13138*, 2021.

[68] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[69] Emmanouil Stergiadis, Satendra Kumar, Fedor Kovalev, and Pavel Levin. Multi-domain adaptation in neural machine translation through multidimensional tagging. *arXiv preprint arXiv:a2102.10160v1*, 2021.

[70] Haoran Devlin, Benjamin Van Durme, and Kenton Murray. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[71] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[72] George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, pages 138–145, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

[73] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, Jun 2005.

[74] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA, August 8-12 2006. Association for Machine Translation in the Americas.

[75] Aaron L. F. Han, Derek F. Wong, and Lidia S. Chao. LEPOR: A robust evaluation metric for machine translation with augmented factors. In *Proceedings of COLING 2012: Posters*, pages 441–450, Mumbai, India, Dec 2012. The COLING 2012 Organizing Committee.

[76] Amjad Almahairi, Kyunghyun Cho, Nizar Habash, and Aaron Courville. First result on arabic neural machine translation. *arXiv preprint arXiv:1606.02680*, 2016.

[77] Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104v4*, 2020.

[78] Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and Stephan Vogel. Qcri machine translation systems for iwslt 16. *arXiv preprint arXiv:1701.03924*, 2017.

[79] Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada, Jul 2017. Association for Computational Linguistics.

[80] Pamela Shapiro and Kevin Duh. Morphological word embeddings for Arabic neural machine translation in low-resource setting. In *Proceedings of the Second Workshop on Subword/Character LEvel Models*, pages 1–11, New Orleans, Jun 2018. Association for Computational Linguistics.

[81] Ebtesam H. Almansor and Ahmed Al-Ani. A hybrid neural machine translation technique for translating low resource languages. In Petra Perner, editor, *Machine Learning and Data Mining in Pattern Recognition*, pages 347–356, Cham, 2018. Springer International Publishing.

[82] Abdullah Alrajeh. A recipe for arabic-english neural machine translation. *arXiv preprint arXiv:1808.06116*, 2018.

[83] Fares Aqlan, Xiaoping Fan, Abdullah Alqwbani, and Akram Al-Mansoub. Arabic–chinese neural machine translation: Romanized arabic as subword unit for arabic-sourced translation. *IEEE Access*, 7:133122–133135, 2019.

[84] Mai Oudah, Amjad Almahairi, and Nizar Habash. The impact of preprocessing on Arabic-English statistical and neural machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 214–221, Dublin, Ireland, Aug 2019. European Association for Machine Translation.

[85] Jezia Zakraoui, Moutaz Saleh, Somaya Al-Maadeed, and Jihad Mohamad AlJa'am. Evaluation of arabic to english machine translation systems. In *2020 11th International Conference on Information and Communication Systems (ICICS)*, pages 185–190, 2020.

[86] Arwa Alqudsi, Nazlia Omar, and Khalid Shaker. Arabic machine translation: a survey. *Artificial Intelligence Review*, 42(4):549–572, 2014.

[87] A Fassi Fehri. *Issues in the structure of Arabic clauses and words*, volume 29. Springer Science & Business Media, 1993.

[88] Achraf Chalabi. Mt-based transparent arabization of the internet tarjim. com. In *Conference of the Association for Machine Translation in the Americas*, pages 189–191. Springer, 2000.

[89] Kevin Daimi. Identifying syntactic ambiguities in single-parse arabic sentence. *Computers and the Humanities*, 35(3):333–349, 2001.

[90] Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[91] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), pages 1715–1725, Berlin, Germany, Aug 2016. Association for Computational Linguistics.

[92] ThuyLinh Nguyen and Stephan Vogel. Context-based Arabic morphological analysis for machine translation. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 135–142, Manchester, England, Aug 2008. Coling 2008 Organizing Committee.

[93] Abraham Ittycheriah and Salim Roukos. A maximum entropy word aligner for Arabic-English machine translation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 89–96, Vancouver, British Columbia, Canada, Oct 2005. Association for Computational Linguistics.

[94] Yasser Salem, Arnold Hensman, and Brian Nolan. Implementing arabic-to-english machine translation using the role and reference grammar linguistic model. In *Proceedings of the Eighth Annual International Conference on Information Technology and Telecommunication*, pages 103–110, 2008.

[95] Omar Shirko, Nazlia Omar, Haslina Arshad, and Mohammed Albared. Machine translation of noun phrases from arabic to english using transfer-based approach. *Journal of Computer Science*, 6(3):350, 2010.

[96] Yassine Benajiba, Imed Zitouni, Mona Diab, and Paolo Rosso. Arabic named entity recognition: Using features extracted from noisy data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 281–285, Uppsala, Sweden, Jul 2010. Association for Computational Linguistics.

[97] Youness Moukafih, Nada Sbihi, Mounir Ghogho, and Kamel Smaïli. Improving machine translation of arabic dialects through multi-task learning. In *20th International Conference Italian Association for Artificial Intelligence:AIxIA 2021*, pages 235–243, MILAN/Virtual, Italy, Dec 2021.

[98] Nouhaila Bensalah, Habib Ayad, Abdellah Adib, and Abdelhamid Ibn El Farouk. Lstm vs. gru for arabic machine translation.

[99] Nouhaila Bensalah, Habib Ayad, Abdellah Adib, and Abdelhamid Ibn El Farouk. Cran: An hybrid cnn-rnn attention-based model for arabic machine translation. In Mohamed Ben Ahmed, Horia-Nicolai L. Teodorescu, Tomader Mazri, Parthasarathy Subashini, and Anouar Abdelhakim Boudhir, editors, *Networking, Intelligent Systems and Security*, pages 87–102, Singapore, 2022. Springer Singapore.

[100] Arianna Bisazza and Marcello Federico. Chunk-based verb reordering in VSO sentences for Arabic-English statistical machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 235–243, Uppsala, Sweden, Jul 2010. Association for Computational Linguistics.

[101] Joseph Aoun, Elabbas Benmamoun, and Dominique Sportiche. Agreement, word order, and conjunction in some varieties of arabic. *Linguistic inquiry*, pages 195–220, 1994.

[102] M Attia. Implications of the agreement features in machine translation. *Al-Azhar University*, 2002.

[103] Ruhi Sarikaya, Yonggang Deng, and Yuqing Gao. Context dependent word modeling for statistical machine translation using part-of-speech tags. In *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[104] Ahmed Hatem and Amin Nassar. Modified dijstra-like search algorithm for english to arabic machine translation system. *Proceedings EAMT*, 2008:12th, 2008.

[105] Kristina Toutanova, Hisami Suzuki, and Achim Ruopp. Applying morphology generation models to machine translation. In *Proceedings of ACL-08: HLT*, pages 514–522, Columbus, Ohio, Jun 2008. Association for Computational Linguistics.

[106] Jakob Elming and Nizar Habash. Syntactic reordering for English-Arabic phrase-based machine translation. In *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages*, pages 69–77, Athens, Greece, Mar 2009. Association for Computational Linguistics.

[107] Ibrahim Badr, Rabih Zbib, and James Glass. Syntactic phrase reordering for English-to-Arabic statistical machine translation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 86–93, Athens, Greece, Mar 2009. Association for Computational Linguistics.

[108] Chafia Mankai and Ali Mili. Machine translation from arabic to english and french. *Information Sciences-Applications*, 3(2):91–109, 1995.

[109] Sasa Hasan, Anas El Isbihani, and Hermann Ney. Creating a large-scale arabic to french statistical machine translation system. In *LREC*, pages 855–858, 2006.

[110] Doaa Samy, Antonio Moreno-Sandoval, José María Guirao, and Enrique Alfonseca. Building a parallel multilingual corpus (arabic-spanish-english). In *LREC*, pages 2176–2181, 2006.

[111] Doaa Samy and Ana González-Ledesma. Pragmatic annotation of discourse markers in a multilingual parallel corpus (arabic-spanish-english). In *LREC*, 2008.

[112] Djamel Mostefa, Mariama Laïb, Stéphane Chaudiron, Khalid Choukri, and G Chalendar. A multilingual named entity corpus for arabic, english and french. *MEDAR*, 2009:2nd, 2009.

[113] Nizar Habash and Jun Hu. Improving Arabic-Chinese statistical machine translation using English as pivot language. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 173–181, Athens, Greece, Mar 2009. Association for Computational Linguistics.

[114] Fathi Debili and Elyes Sammouda. Aligning sentences in bilingual texts French - English and French - Arabic. In *COLING 1992 Volume 2: The 15th International Conference on Computational Linguistics*, 1992.

[115] Marine Carpuat, Yuval Marton, and Nizar Habash. Improving Arabic-to-English statistical machine translation by reordering post-verbal subjects for alignment. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 178–183, Uppsala, Sweden, Jul 2010. Association for Computational Linguistics.

[116] Chadia Moghrabi. On parametering the choice of words in text generation and its usefulness in machine translation. In *Machine translation: ten years on*, pages 1–9, Cranfield University, England,, Nov 1998. Cranfield University Press.

[117] Mathieu Guidere. Toward corpus-based machine translation for standard arabic. *Translation Journal*, 6(1), 2002.

[118] Haytham Alsharaf, Sylviane Cardey, and Peter Greenfield. French to arabic machine translation: the specificity of language couples. In *Proc. of the 9th Annual Workshop of the European Association for Machine Translation (EAMT), Malta, April*, 2004.

[119] Mark Pedersen, Domenyk Eades, Samir K. Amin, and Lakshmi Prakash. Relative clauses in Hindi and Arabic: A paninian dependency grammar analysis. In *Proceedings of the Workshop on Recent Advances in Dependency Grammar*, pages 9–16, Geneva, Switzerland, Aug 28 2004. COLING.

[120] Pierrette Bouillon, Ismahene Sonia Halimi Mallem, Yukie Nakao, Kyoko Kanzaki, Hitoshi Isahara, Nikolaos Tsourakis, Marianne Starlander, Beth Ann Hockey, and Emmanuel Rayner. Developing non-european translation pairs in a medium-vocabulary

medical speech translation system. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, pages 1741–1748, 2008.

[121] Romaric Besançon, Djamel Mostefa, Ismaïl Timimi, Stéphane Chaudiron, Mariama Laïb, and Khalid Choukri. Arabic, english and french: three languages in a filtering systems evaluation project. *MEDAR*, pages 163–167, 2009.

[122] Ahmad T Al-Taani and Zeyad M Hailat. A direct english-arabic machine translation system. *Information Technology Journal*, 4(3):256–261, 2005.

[123] Abraham Ittycheriah and Salim Roukos. Direct translation model 2. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 57–64, Rochester, New York, Apr 2007. Association for Computational Linguistics.

[124] Mouiad Alawneh and Tengku Mohd Sembok. Handling agreement and words reordering in machine translation from english to arabic using hybrid-based systems. *Journal of Computer Science*, 11(6):93–97, 2011.

[125] Khaled Shaalan. Rule-based approach in arabic natural language processing. *The International Journal on Information and Communication Technologies (IJICT)*, 3(3):11–19, 2010.

[126] Kfir Bar and N. Dershowitz. Using verb paraphrases for arabic-to-english example-based translation. *Machine Translation and Morphologically-rich Languages*, 2011.

[127] Kfir Bar and N. Dershowitz. *Using semantic equivalents for Arabic-to-English: Example-based translation*, pages 49–72. John Benjamins Publishing, Amsterdam, Netherlands, 2012.

[128] A Farghali. Arabic machine translation: A developmental perspective. *International Journal on Information and Communication Technologies*, 3(3), 2010.

[129] Rached Zantout and Ahmed Guessoum. Arabic machine translation: A strategic choice for the arab world. *Journal of King Saud University - Computer and Information Sciences*, 12:117–144, Dec 2000.

[130] Violetta Cavalli-Sforza and Aaron Phillips. *Using morphology to improve Example-Based Machine Translation*, pages 23–48. John Benjamins Publishing, Jan 2012.

[131] T El-Shishtawy and A El-Sammak. The best templates match technique for example based machine translation. *arXiv preprint arXiv:1406.1241*, 2014.

[132] Evgeny Matusov, Gregor Leusch, and Hermann Ney. *Learning To Combine Machine Translation Systems*, chapter 13, pages 257–276. Neural Information Processing Series. MIT Press, 2009.

[133] Emad Mohamed and Fatiha Sadat. Hybrid arabic–french machine translation using syntactic re-ordering and morphological pre-processing. *Computer Speech & Language*, 32(1):135–144, 2015.

[134] Manar Alkhatib and Khaled Shaalan. Paraphrasing arabic metaphor with neural machine translation. *Procedia Computer Science*, 142:308–314, 2018.

[135] Sara Ebrahim, Doaa Hegazy, Mostafa Gadal Haqq M. Mostafa, and Samhaa R. El-Beltagy. Detecting and integrating multiword expression into english-arabic statistical machine translation. *Procedia Computer Science*, 117:111–118, 2017.

[136] Nizar Habash, B. Dorr, and Christof Monz. Symbolic-to-statistical hybridization: extending generation-heavy machine translation. *Machine Translation*, 23:23–63, 2009.

[137] Nizar Y. Habash. Introduction to arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1):1–187, 2010.

[138] Laith H Baniata, Seyoung Park, and Seong-bae Park. A neural machine translation model for arabic dialects that utilizes multitask learning (mtl). *Computational Intelligence and Neuroscience*, 2018:1–10, 2018.

[139] Fatma Mallek, Billal Belainine, and Fatiha Sadat. Arabic social media analysis and translation. *Procedia Computer Science*, 117:298–303, 2017.

[140] Nizar Habash, Nasser Zalmout, Dima Taji, Hieu Hoang, and Maverick Alzate. Parallel corpus for evaluating machine translation between Arabic and European languages. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 235–241, Valencia, Spain, apr 2017. Association for Computational Linguistics.

[141] Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054—-2059, Barcelona, Spain (online), 2020. International Committee for Computational Linguistics.

[142] Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. Incorporating BERT into Neural Machine Translation. In *Proceedings*

of the Eighth International Conference on Learning Representations, Addis Abbaba, Ethiopia (Online), 2020.

[143] Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. The interplay of variant, size, and task type in arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Online), 2021. Association for Computational Linguistics.

[144] Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samiha. Pretraining bert on arabic tweets: Practical considerations. *arXiv Preprint:2102.10684*, 2021.

[145] Marwa Gaser, Manuel Mager, Injy Hamed, Nizar Habash, Slim Abdennadher, and Ngoc Thang Vu. Exploring segmentation approaches for neural machine translation of code-switched egyptian arabic-english text, 2022.

[146] Aissam Outchakoucht and Hamza Es-Samaali. Moroccan dialect-darija-open dataset. *arXiv preprint arXiv:2103.09687*, 2021.

[147] Gyu-Hyeon Choi, Jong-Hun Shin, and Young-Kil Kim. Improving a multi-source neural machine translation model with corpus extension for low-resource languages. *arXiv Preprint:1709.08898*, 2017.

[148] Mohamed Seghir Hadj Ameur, Ahmed Guessoum, and Farid Meziane. Improving arabic neural machine translation via n-best list re-ranking. *Machine Translation*, 33(4):279–314, 2019.

[149] Safae Berrichi and Azzeddine Mazroui. Addressing limited vocabulary and long sentences constraints in english–arabic neural machine translation. *Arabian Journal for Science and Engineering*, 46(9):8245–8259, 2021.

[150] Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China, nov 2019. Association for Computational Linguistics.

[151] Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference*

on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, may 2018. European Language Resources Association (ELRA).

[152] Chenchen Ding, Hnin Thu Zar Aye, Win Pa Pa, Khin Thandar Nwet, Khin Mar Soe, Masao Utiyama, and Eiichiro Sumita. Towards burmese (myanmar) morphological analysis: Syllable-based tokenization and part-of-speech tagging. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 19(1), may 2019.

[153] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020.

[154] Ibrahim Gashaw and HL Shashirekha. Amharic-arabic neural machine translation. *arXiv preprint arXiv:1912.13161*, 2019.

[155] Stéphane Clinchant, Kweon Woo Jung, and Vassilina Nikoulina. On the use of bert for neural machine translation. *arXiv Preprint:1909.12744*, 2019.

[156] Wuwei Lan, Yang Chen, Wei Xu, and Alan Ritter. An empirical study of pre-trained transformers for arabic information extraction. In *Proceedings of The 2020 Conference on Empirical Methods on Natural Language Processing (EMNLP)*, 2020.

[157] Amel Slim, Ahlem Melouah, Usef Faghihi, and Khouloud Sahib. Improving neural machine translation for low resource algerian dialect by transductive transfer learning strategy. *Arabian Journal for Science and Engineering*, pages 1–8, 2022.

[158] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *Proceedings of the Eighth International Conference on Learning Representations*, Addis Abbaba, Ethiopia (Online), 2020.

[159] Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. Arbert and marbert: Deep bidirectional transformers for arabic. *arXiv Preprint:2101.01785*, 2020.

[160] Rasha Al Dam and Ahmed Guessoum. Building a neural network-based english-to-arabic transfer module from an unrestricted domain. In *2010 International Conference on Machine and Web Intelligence*, pages 94–101, 2010.

[161] RACHED Zantout and Ahmed Guessoum. Obstacles facing arabic machine translation: building a neural network-based transfer module. *Papers in Translation Studies*, pages 229–253, 2015.

[162] Young-Suk Lee, Kishore Papineni, Salim Roukos, Ossama Emam, and Hany Hassan. Language model based Arabic word segmentation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 399–406, Sapporo, Japan, Jul 2003. Association for Computational Linguistics.

[163] Paisarn Charoenpornsawat, Virach Sornlertlamvanich, and Thatsanee Charoenporn. Improving translation quality of rule-based machine translation. In *COLING-02: Machine Translation in Asia*, 2002.

[164] Marwan Akeel and Ravi Mishra. Ann and rule based method for english to arabic machine translation. *Int. Arab J. Inf. Technol.*, 11(4):396–405, 2014.

[165] Mossa Ghurab, Yueting Zhuang, Jiangqin Wu, and Maan Younis Abdullah. Arabic-chinese and chinese-arabic phrase-based statistical machine translation systems. *Inf. Technol. J*, 9(4):666–672, 2010.

[166] Fares Aqlan, Xiaoping Fan, Abdullah Alqwbani, and Akram Al-Mansoub. Improved arabic-chinese machine translation with linguistic input features. *Future Internet*, 11:22, 2019.

[167] Khaled Shaalan and Ahmad Hany Hossny. Automatic rule induction in arabic to english machine translation framework. *Challenges for Arabic Machine Translation*, 9(2012):135, 2012.

[168] Khaled Shaalan, Ashraf Hendam, and Ahmed Rafea. An English-Arabic bi-directional machine translation tool in the agriculture domain: A rule-based transfer approach for translating expert systems. *IFIP Advances in Information and Communication Technology*, pages 281–290, 2010.

[169] Arwa Hatem, Nazlia Omar, and Khalid Shaker. Morphological analysis for rule based machine translation. In *2011 International Conference on Semantic Technology and Information Retrieval*, pages 260–263. IEEE, 2011.

[170] Sameh Alansary. Interlingua-based machine translation systems: Unl versus other interlinguas. *The Egyptian Journal of Language Engineering*, 1, Jan 2014.

[171] Khaled Shaalan, Azza Abdel Monem, Ahmed Rafea, and Hoda Baraka. Mapping interlingua representations to feature structures of arabic sentences. In *The Challenge of Arabic for NLP/MT. International Conference at the British Computer Society, London*, pages 149–159, 2006.

[172] Khaled Shaalan, Ahmed Rafea, Azza Abdel Moneim, and Hoda Baraka. Machine translation of english noun phrases into arabic. *International Journal of Computer Processing of Oriental Languages*, 17(02):121–134, 2004.

[173] M. M. Abu Shquier and T. M. T. Sembok. Word agreement and ordering in english-arabic machine translation. In *2008 International Symposium on Information Technology*, volume 1, pages 1–10, 2008.

[174] Mona Diab, Mahmoud Ghoneim, and Nizar Habash. Arabic diacritization in the context of statistical machine translation. In *Proceedings of MT-Summit*, 2007.

[175] Andrea Zaninello and Alexandra Birch. Multiword expression aware neural machine translation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3816–3825, Marseille, France, May 2020. European Language Resources Association.

[176] Valia Kordoni and Iliana Simova. Multiword expressions in machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1208–1211, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[177] Kfir Bar, Mona Diab, and Abdelati Hawwari. *Arabic Multiword Expressions*, pages 64–81. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014.

[178] Kenneth R. Beesley. Arabic finite-state morphological analysis and generation. In *Proceedings of the 16th conference on association for computational linguistics*, pages 89–94, Jan 1996.

[179] Everhard Ditters. A formal grammar for the description of sentence structure in modern standard arabic. In *In EACL 2001 Workshop Proceedings on Arabic Language Processing: Status and Prospects*, pages 31–37, 2001.

[180] Alexandre Rafalovitch and Robert Dale. United Nations general assembly resolutions: A six-language parallel corpus. In *Proceedings of Machine Translation Summit XII: Posters*, Ottawa, Canada, aug 26-30 2009.

[181] Faical Azouaou and Imane Guellil. Alg/fr: A step by step construction of a lexicon between algerian dialect and french. In *The 31st Pacific Asia Conference on Language, Information and Computation PACLIC*, volume 31, 2017.

[182] Jörg Tiedmann. Parallel data, toolsand interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, volume 2012, pages 2214–2218, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[183] Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 5960–5969, Online, November 2020. Association for Computational Linguistics.

[184] Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500. Association for Computational Linguistics, aug 2021.

[185] Pierre Lison and Jörg Tiedemann. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia, may 2016. European Language Resources Association (ELRA).

[186] Hamdy Mubarak. Dial2MSA: A tweets corpus for converting dialectal arabic to modern standard arabic. In *OSACT 3: The 3rd Workshop on Open-Source Arabic Corpora and Processing Tools*, page 49, 2018.

[187] Kheireddine Abainia. DZDC12: a new multipurpose parallel Algerian Arabizi–French code-switched corpus. *Language Resources and Evaluation*, 54(2):419–455, jun 2020.

[188] Mustafa Jarrar, Nizar Habash, Diyam Akra, and Nasser Zalmout. Building a corpus for palestinian arabic: a preliminary study. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP)*, pages 18–27, Doha, Qatar, oct 2014.

[189] Karima Meftouh, Salima Harrat, Salma Jamoussi, Mourad Abbas, and Kamel Smaili. Machine translation experiments on PADIC: A Parallel Arabic DIalect Corpus. In

Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, pages 26–34, Shanghai, China, oct 2015.

[190] Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, may 2018. European Language Resources Association (ELRA).

[191] Stephen Grimes, Xuansong Li, Ann Bies, Seth Kulick, Xiaoyi Ma, and Stephanie Strassel. Creating arabic-english parallel word-aligned treebank corpora at ldc. In *Proceedings of Language Resources and Evaluation Conference (LREC'10), Malta*, 2010.

[192] Hassan Sawaf. Arabic dialect handling in hybrid machine translation. In *AMTA 2010 - 9th Conference of the Association for Machine Translation in the Americas*, 2010.

[193] Salam Khalifa, Nizar Habash, Dana Abdulrahim, and Sara Hassan. A large scale corpus of Gulf Arabic. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4282–4289, Portorož, Slovenia, may 2016. European Language Resources Association (ELRA).

[194] Ann Bies, Zhiyi Song, Mohamed Maamouri, Stephen Grimes, Haejoong Lee, Jonathan Wright, Stephanie Strassel, Nizar Habash, Ramy Eskander, and Owen Rambow. Transliteration of Arabizi into Arabic orthography: Developing a parallel annotated Arabizi-Arabic script SMS/chat corpus. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 93–103, Doha, Qatar, oct 2014. Association for Computational Linguistics.

[195] Spence Green and Christopher D. Manning. Better Arabic parsing: Baselines, evaluations, and analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 394–402, Beijing, China, aug 2010. Coling 2010 Organizing Committee.

[196] Selçuk Köprü and Jude Miller. A unification based approach to the morphological analysis and generation of arabic. In *CAASL3: Proc. of the 3rd Workshop on Computational Approaches to Arabic-script based Languages*, Ottawa, ON, Canada, 2009.

[197] Sameh Alansary, Magdy Nagi, and Noha Adly. Towards analyzing the international corpus of arabic (ica): Progress of morphological stage. In *8th International Conference on Language Engineering, Egypt*, pages 1–23, 2008.

[198] Mohammed Attia. *Handling Arabic morphological and syntactic ambiguity within the LFG framework with a view to machine translation.* The University of Manchester (United Kingdom), 2008.

[199] Mohammed Attia. Developing a robust arabic morphological transducer using finite state technology. In *8th annual CLUK research colloquium*, pages 9–18. Citeseer, 2005.

[200] Eman Othman, Khaled Shaalan, and Ahmed Rafea. A chart parser for analyzing modern standard Arabic sentence. In *Workshop on Machine Translation for Semitic languages: issues and approaches*, New Orleans, USA, sep 23-27 2003.

[201] Zdeněk Žabokrtský and Otakar Smrž. Arabic syntactic trees: from constituency to dependency. In *10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary, apr 2003. Association for Computational Linguistics.

[202] Kareem Darwish. Building a shallow Arabic morphological analyser in one day. In *Proceedings of the ACL-02 Workshop on Computational Approaches to Semitic Languages*, Philadelphia, Pennsylvania, USA, jul 2002. Association for Computational Linguistics.

[203] Leah S. Larkey, Lisa Ballesteros, and Margaret E. Connell. Improving stemming for arabic information retrieval: Light stemming and co-occurrence analysis. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, page 275–282, New York, NY, USA, 2002. Association for Computing Machinery.

Appendix A

Title of Appendix A

Here goes the appendix A

2nd page of appendix A

Appendix B

Title of Appendix B

Here goes the appendix B

2nd page of appendix B