Toward a Test Set of Dislocations in Persian for Neural Machine Translation

Behnoosh Namdarzadeh[1], Nicolas Ballier[1 2] , Lichao Zhu[1], Guillaume Wisniewski[2], Jean-Baptiste Yunès[3]
[1]CLILLAC-ARP, [2]LLF, [3]IRIF
Université Paris Cité, F-75013 Paris, France
behnoosh.namdarzadeh@etu.u-paris.fr,
{nicolas.ballier, lichao.zhu, guillaume.wisniewski, jean-baptiste.yunes}@u-paris.fr

Abstract

This paper describes a test set designed to analyse the translation of dislocations from Persian, to be used for testing neural machine translation models. We first tested the accuracy of the two Universal dependency treebanks for Persian to automatically detect dislocations. Then we parsed the available Persian treebanks on GREW (Bonfante et al., 2018) to build a specific test set containing examples of dislocations. With available aligned data on OPUS (Tiedemann, 2016), we trained a model to translate from Persian into English on openNMT (Klein et al., 2017). We report the results of our translation test set by several toolkits (Google Translate, MBART-50 (Tang et al., 2020), Microsoft Bing and our in-house translation model) for the translation into English. We discuss why dislocations in Persian provide an interesting testbed for neural machine translation.

1 Introduction

This paper describes a first experiment (to the best of our knowledge) at building a neural machine translation test corpus relying on Persian dislocations. Dislocation is a structure that allows the repetition of a dislocated item with (usually) a proform that resumes the referent of the dislocated item like من، خودم... 'English: *I, myself ...*' 'French: *moi, je...*'. Previous research has shown that dislocations can be challenging for neural machine translation, because they tend to be very present in spoken data and consequently often under-represented in training data, resulting in mistranslations, for example from French into English where the dislocated item is often reduplicated with a second agrammatical subject (Namdarzadeh and Ballier, 2022). For neural machine translation (NMT), dislocations are therefore challenging and a perfect topic for a challenge set approach (Isabelle et al., 2017).

Persian still is as an under-resourced language for NLP tasks, as shown in the Proceedings of the NSURL Workshop (Freihat and Abbas, 2021). From a typological perspective, not only does Persian allow dislocation like many other languages, but also scrambling (?), so that investigating the translation of dislocated constructions raises interesting linguistic questions in the direction of fixed ordered languages like English. Our combination of languages is an interesting observatory to investigate the translation of word order. Two main research questions are addressed: do we observe an agrammatical copy of the dislocated item in the translation (syntactic adequacy) and is the information packaging effect of the dislocation rendered in the translation (pragmatic adequacy)?

The rest of the paper is structured as follows: Section 2 provides an overview of Machine Translation (MT) related resources for Persian. Section 3 explains how we collected the dislocations from existing Treebanks. Section 4 describes the translation model we produced. Section 5 analyses the translations produced by different MT systems we tested. Section 6 discusses our findings.

2 Previous Research and Resources

Persian, also known as Farsi, is an Indo-Iranian branch of the Indo-European family. Persian has three variants: Western Persian, referred to as 'Parsi' or 'Farsi' which is spoken in Iran. Eastern Persian referred to as 'Dari' and spoken in Afghanistan. And the last variant is Tajiki, which is spoken in Tajikistan and Uzbekistan

(Seraji, 2015).

2.1 Previous MT systems

One of the prototype translation systems that is able to translate Persian into English is the Shiraz machine translation project (Amtrup et al., 2000). Feeding the translation model with the higher size of parallel corpora from different domains improved the outputs of the system significantly (Mohaghegh, 2012). Years later, the emergence of MIZAN corpus, the biggest Persian-English parallel corpus, can be considered as an improvement in the field of machine translation. It consists of 1,021,596 Persian-English aligned sentences. An SMT system was developed using this corpus to observe the function of the translation model. Despite the acceptable BLEU score, the conclusion is that Persian remains an under-resourced language with comprehensive open issues (Kashefi, 2018).

2.2 Previous NMT systems

For neural machine translation (NMT), Persian is not (as yet?) implemented in DeepL but in Google Translate toolkit and no less than 14 APIs support Persian for MT [1]. Several dictionaries for English to Farsi are available online [2]. We resorted to the online versions of Google Translate, Bing Microsoft Translator (hereafter Bing) and MBART-50, the multilingual model developed for 50 languages (Tang et al., 2020).

2.3 Available UD Treebanks for Persian

For the analysis of Persian using Universal dependency (De Marneffe et al., 2006; De Marneffe and Manning, 2008), two treebanks have been developed: (Seraji et al., 2016) and (Rasooli et al., 2020) deriving from the Persian Dependency Treebank (Rasooli et al., 2013). We searched for examples of dislocations in the treebanks and report our findings in the following section.

3 Dislocations in Persian

3.1 Previous Research

Before beginning the typologies of plausible dislocated constructions in Persian, we have to pinpoint that Persian is a pro-drop language. This means that the agreement between the verb and its subject is realized by verbal suffixes (Faghiri and Samvelian, 2021); thus the subject can be dropped in a sentence. Persian displays free word order (Faghiri and Samvelian, 2021) but is an SOV language. There are some cases in Persian where the SOV canonical word ordering is changed based on the context. This can be clearly seen in a sentence where the constituent گل'flower' is positioned at the left side of the sentence, expressing the contrastive focus in the sentence گل علی برای مریم خرید 'flower Ali for Maryam buy-PST' that the subject buys گل 'flower' and not something else (Faghiri and Samvelian, 2021). The other dislocated element in Persian is quite similar to the French *ce que* structure. In the Persian sentence آنچه که گفت درست بود the sentence begins with آنچه (what), meaning *What (s)he said was right* (Faghiri and Samvelian, 2021). Furthermore, clefting is frequent in Persian, in a way that the focused element is moved to the initial position of a sentence. Various functions can be cloven except adverbs, like in the example توی باغ بود که همدیگر را دیدیم 'in garden be-PST that each other ra see-PL' the adjunct is cloven (Faghiri and Samvelian, 2021).

3.2 Data Collection with `GREW`

We also queried the UD_PersianSeraji treebank on the `GREW` project[3]. Figure 1 shows the "relation table" (Guibon et al., 2020) which displays the relations between a governor (here, selected with the category "dislocated") and the corresponding dependents, classified as columns according to their part of speech (here, nouns, pronouns and particles).

It can be also argued that manipulating some of the examples, placing the تو خودت in the left periphery of the sentence changes the detection of the constituent as `dislocated`. It seems that the number of words between the dislocated item and the constituent resumed by the constituent affects the detection of `dislocated`. Interestingly, in the examples taken from `GREW`, اینجا برنامه برای ارتباط با مخاطب خودش دچار مشکل می‌شود., there is a distance between the dislocated item برنامه and خودش, this is not recognized as a dislo-

[1] `https://machinetranslate.org/persian`

[2] e.g. `https://translate.glosbe.com/en-fa/machine%20translation`

[3] `http://universal.grew.fr/?corpus=UD_Persian-Seraji@2.10`

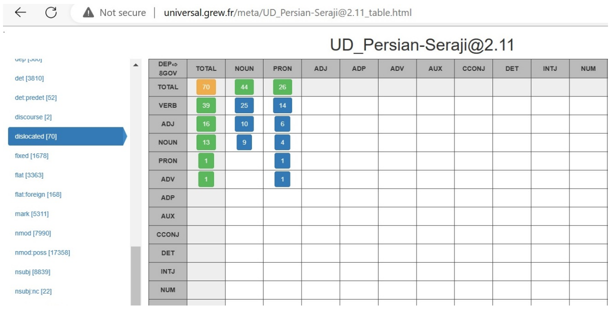

Figure 1: Distribution of Dependent items using GREW

cation item in UDPipe, whereas, in the می‌گویند چه جوری است که تهیه‌کننده خودش دارد تقلب می‌کند!, the dislocated item تهیه کننده and its resuming construction خودش is placed one after another with no distance. The {UDPipe} package (Wijffels, 2022) in R (R Core Team, 2022) can correctly detect it as a dislocated construction. Thus, it might be the case that the proximity of the proposed dislocated constituent to its referent could have an impact on their detection.

3.3 The Two UD Treebanks for Persian

Two Treebanks are currently available for Universal Dependency on github : Persian-Seraji and UD Persian. There are only two Treebanks available in the Universal Dependency (UD) framework. This can be a good reason to label Persian as an under-resourced language. One is PerUDT (Rasooli et al., 2013), which consists of 29,000 sentences extracted from contemporary Persian texts in different genres such as news, academic papers, articles and fictions. The other is UPDT Treebank (Seraji et al., 2016), which consists of 6,000 annotated and validated sentences of different genres. The `GREW-match` project also represents an analysis of the two above-mentioned treebanks in more details. It so happens that dislocations is a hapax in the reference Persian Dependency Treebank (Rasooli et al., 2013). The treebank contains 29,107 sentences and only one occurrence of 'dislocated' was spotted. For the purpose of this study, since no `dislocated` was found in PerUDT Treebank, we chose the UPDT Treebank. We review the dependency relations on `GREW-match` as well, to recheck the annotations and compile the Persian sentences with a `dislocated` dependency relation (deprel).

4 Material and Methods

This section describes how we built the neural translation engine we produced.

4.1 Tokenizations

We used BPE to tokenize English and Persian data sets into subwords by processing as follows: i) first word tokenization of datasets (train, dev, test) is applied with a standard tokenizer of each language; ii) training of a subword tokenization model with monolingual data; iii) a second subword tokenization is applied to the tokenized datasets; iv) training of our neural model with subword-tokenized English↔Persian parallel corpus.

To try to avoid subtokenisation issues, we trained our BPE model with a larger corpus. The data sets for the BPE model are split as follows: for English, we used `spaCy` (Honnibal and Johnson, 2015) library to tokenize a data set, by normalizing and compiling WMT15's Europarl, News Commentary and Common Crawl (Bojar et al., 2015) French↔English parallel corpus, which contains 116,035,319 words. The compiled data set was used to train a `SentencePiece` (Kudo, 2018) BPE model as follows : vocab-size=32000, character_coverage=1, model_type=unigram. As for Persian, we used `Stanza` (Qi et al., 2020) with the UD Persian Seraji Treebank (Qi et al., 2018) to tokenize a Farsi data set (98,472,761 words) from the CCAligned v1 corpus (El-Kishky et al., 2020), in order to train a `SentencePiece` BPE model with comparable data size and with the following parameters : vocab-size=32000, character_coverage=0.9995, model_type=unigram.

4.2 Training

We used TED2020 (Reimers and Gurevych, 2020) Farsi↔English parallel corpus (EN : 6,036,185 words, FA : 7,362,765 words) to train a neural machine translation model with `OpenNMT` (Klein et al., 2017). Both Farsi and English corpora are split into three data sets : `dev` (2,000 lines), `test` (2,000 lines) and `train` (the rest of the data set). `OpenNMT` implements a `transformer` model with the following architecture: 6 encoder and decoder layers; each layer has 8 attention heads; the feed-forward layers of the transformer have 2,046 parame-

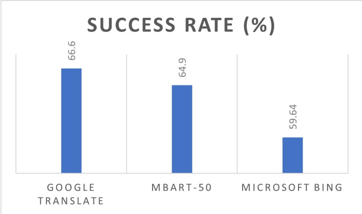

Figure 2: Success Rate of `dislocated` constituent translation

ters; the dimension of word embedding is 512. In the end, there are 72,924,862 parameters.

5 Results

5.1 Detection of Dislocation on Current UD Models

When parsing our test set with the {UDpipe}, only 17 cases out of 57 sentences were detected as dislocation. This means that only about 30% of our examples are recognized as a dislocated item in the test set. Interestingly, there seems to be an identifiable pattern through which dislocated dependency relations are identified. To give a concrete example, duplicated use of the subject is detected by {UDpipe} when singular (e.g. من for *I* or تو for *you*) but not so if plural. Yet, this is not even systematic.

5.2 Quality Evaluation of the Translations across NMT Toolkits

For the evaluation of the quality of the translations, we applied the "descriptive-comparative human analysis" model of Keshavarz , which suggests different types of errors in the outputs, to evaluate the translations (Zand Rahimi et al., 2017). What matters in our evaluation of the quality of the translations is the grammaticality of the translations. Our success rate is based on syntactic adequacy, i.e. avoiding copying the dislocated items in outputs. Compared with more elaborate criteria of human assessment methods (HA) which also analyse fluency and fidelity (Han et al., 2021), we mostly focused on (syntactic) adequacy and comprehension of the outputs rather than on subtle analyses of semantic and pragmatic adequacy. Figure 2 globally indicates success rates of the `dislocated` constituent translations across the different toolkits. Dislocation remains an issue for at least a third of our 57 examples. Among the three toolkits, Google Translate records the highest success rate (66.6 %), and Microsoft Bing gets the lowest rate (59.64%). MBART-50 is in between in this regard (64.9 %, no significant difference, p-value: > 0.05). The individual performance of the three toolkits are discussed in the following subsections.

5.2.1 Translation of Dislocations by Google

Overall, Google outputs tend to follow the English canonical word order, where the (initial) dislocated item in Persian tends to be translated in its expected canonical position in English. Nevertheless, compared to other toolkits, Google Translate uses more dislocated constituents in its output, especially for reflexive dislocated constituents. Out of the 31 cases of reflexive pronouns, 17 were translated following the Persian word order. For example, من خودم در اصفهان هستم 'I-1st-sg self-1st-sg in Esfahân be-v-pre-1st-sg' has the personal pronouns translated as 'I myself [am in Isfahan]' . We do not have access to Google's training data, but checking the COCA (Corpus of Contemporary American English) and the BNC (British National Corpus), we suggest that the toolkit has a translation which is consistent with observed frequencies, at least in the American English reference corpus: *I myself am* occurs 375 and 15 times, and *I am myself* occurs 125 and 18 times in COCA and BNC, respectively. This may hint that American English might be more present than British English in the training data.

5.2.2 Translation of Dislocations by MBart-50

What is observed in the outputs of MBart-50 is similar to what we have seen in Google Translate. Being closer to the English word order than to the Persian word order may lead to over-translation and sometimes to an incorrect rendering of the source sentence. Some of the examples of dislocations in our data exhibit re-arranging to the English canonical order constituents that are "scrambled" in Persian. Analysing the outputs of MBart-50, we might say that the translation engine does not take into consideration this property of Persian (scrambling), tending to translate sentences strictly following the English word order. Like in this example, گرما رو ازش متنفرم 'heat-râ from-

3sg hate-1sg', the MBart-50 translation *I hate the heat* has 'heat' positioned as object, in its standard SVO position, whereas we may expect 'As for the heat, I hate it' (Azizian et al., 2015) . Topicalization of the object intends to focus addressees' attention on this constituent in the Persian sentence, and the translation by MBart-50 disregards this phenomenon, sticking to the standard word ordering. We could say that the NMT outputs meet syntactic adequacy but not exactly pragmatic adequacy.

5.2.3 Translation of Dislocations by Microsoft Bing

Microsoft Bing records the lowest success rate among our toolkits. This means that it tends to copy the `dislocated` constituents, and it also tries to stick to the English canonical word order. The output for the above-mentioned example م` خودم در اصفهان هستم 'I-1st-sg self-1st-sg in Esfahân be-v-pre-1st-sg' is *I am in Isfahan myself*. Again, the presence of *myself* in final position is frequent in reference corpora (40,265 and 2,141 occurrences in COCA and BNC, respectively, with a high Log likelihood for the American data, 984.96).

Compared to other toolkits, on our (limited) set of examples, Bing produces more nonsense translations for English. In some cases, the very meaning of the source text is ruined. For example, translating the Persian sentence کتابو سامان فرستاد'book-Obj Râ Saman-Sbj send-3sg-pst' (possible translation: *The book, Saman sent.*), Microsoft Bing entirely deteriorates what was said in the source text by the output *Saman's book sent him*. The example clearly indicates that topicalized noun phrase and copy of the same subject in the source sentence can be challenging for the current state of the translation model.

5.2.4 Translation Produced by our Prototype Model

Our translations were far from satisfactory, probably due to data scarcity of training data, though MBART-50 uses only a selection (and a filtered selection) of the TED talk data we used[4]. For MBART-50, they used (after filtering) 14,4895 sentences from TED58 for train, 3,930 for validation and 4,490 sentences for test according to the Appendix of (Tang et al., 2020). Additional data building on Perlex (Asgari-Bidhendi et al., 2021) or exploiting the monolingual BERT for the Persian language (ParsBERT) (Farahani et al., 2021) might be a way to improve the performance of our system.

6 Discussion

6.1 Scrambling and Translations in Fixed Order Languages

Analysing dislocations offers a bird's eye view on a crucial typological distinction between Persian and English. If English has a fixed word order, Persian like some other languages, allows "scrambling", i.e. it has the ability to change word order without changing the meaning (Ross, 1967). The research question can be reformulated, from the point of view of Persian, as "should we pragmatically expect a non-canonical order in the translation?" More generally, does the translation of languages that allow scrambling require a specific word order, for example exploiting Left Periphery? For argument's sake, we investigated the translation of dislocations by MBart-50 into French, which potentially has dislocations, especially in its left periphery. Since French was also included in the 50 languages and is famous for its dislocations, we analysed the outputs in French to see if dislocations were used in the French translations. The copied structures from Persian are not transferred into French in most of the cases. The Persian possessive pronouns are not conveyed in French, and in some other cases, the French output does not make sense, indicating a deficiency in the training process. Hallucinations (Raunak et al., 2021) where outputs are barely related to their source texts can be observed as well as English words in the French translations.

6.2 Pragmatic Adequacy or just Syntactic Adequacy?

Investigating word order in the translation leads us to a more surface analysis of constituents (syntactic adequacy, meeting the requirements of the canonical word order) but paying attention to the possible modifications

[4]To verify our hypothesis, we have trained a second OpenNMT `transformer` model following the same process, by using CCAligned fa↔en parallel corpus as training data, which are 10 times larger than TedTalk corpus. The translations produced by the model are much more relevant.

of the word order leads to a more semantic/pragmatic perspective. Linear arrangement of linguistic elements in a sentence has a role in ”processing information and organizing messages at text level” (Baker, 2011). Especially when it comes to spoken data, information structure can be even more complex to capture and interpret. Thus, taking into consideration the *information packaging* of the sentence, including ”syntactic, prosodic, and morphological means” plays a crucial role (Vallduví and Engdahl, 1996). Within a text linguistic approach, the clause position is posited as containing a discourse-pragmatic function cross-linguistically. To give an example, the peripheral modifiers in the clause in Persian are placed relatively freely and indicate different discourse functions. In other words, the placement of main and peripheral constituents within a sentence is more determined by semantic and pragmatic factors than by solid rules. In contrast, English syntactic structures are controlled by the grammatical rules. For instance, the constituent that precedes the verb must be subject and the verb must be immediately followed by a direct object (Roberts et al., 2009).

Depending on the position of dislocated constituents within a sentence, we may understand that the speaker tries to introduce a new topic or uses this linguistic device to indicate a contrastive focus. The dislocated constituent might also be used to re-state a given topic for discourse cohesion (Karimi, 2005). We might discuss whether dislocated constructions in the source text should remain a scrambled segment in the target text.

7 Conclusion

In this paper, we have described some existing NLP resources for Persian in relation to Neural Machine Translation. We described how we built our test set extracting examples with the `dislocated` dependency relation from Persian universal dependency treebanks on `GREW`.[5] Though limited in size, it showed issues in more than a third of the translations produced by Google Translate, MBart-50 and Microsoft Bing. The answer to our first research question (do we observe an agrammatical copy of the dislocated item in the translation?) is negative. Our conclusion is that toolkits tend to preserve the canonical structure of an English sentence when it comes to translating Persian dislocated items and topicalized constituents. This partially answers our second research question : the information packaging effect of the dislocation is only partially rendered in the translations.

What is crucial in this challenge set based study is to come up with a challenging structure that is used to probe the NMT toolkits. Dislocation seems a challenging one, since this is not a frequent structure in English. The very question we might ask ourselves is to what extent we expect the system to preserve a dislocated segment in its output. Based on what we have seen in the translations from Persian into English, when the doubled structure does not capture in the translation, the core meaning of the sentence changes. Using a “scrambled” sentence with non-canonical word order, the speaker has a certain purpose. Translating it into the canonical order might ruin the very purpose of the speaker and might not convey the exact state-of-affairs in discourse. Thus, to reach pragmatic adequacy, it might be suggested that the dislocated item in Persian be given a specific status in information structure in the target sentence. It might be excessive to suggest that we should expect the systems to produce a sentence preserving the non-canonical structure of the source text. Since dislocations are mostly used in spoken data, we can suggest that systems are probably not sufficiently trained with this type of data. In this sense, to align with frequent structures in spoken data, our challenge set could be expanded using other grammatical phenomena such as *it*-clefts and pseudo-clefts or to include cases of local scrambling and long distance scrambling (Rezaei, 2000).

Acknowledgments

This publication has emanated from research supported in part by the 2020 “émergence” research project SPECTRANS, under the ANR grant (ANR-18-IDEX-0001, Financement IdEx Université Paris Cité).

[5] We completed our test set of 57 examples, to be found on `https://github.com/nballier/SPECTRANS/tree/main/NSUR` with examples from (Yousef and Torabi, 2021) and (Azizian et al., 2015).

References

Jan W Amtrup, Hamid Mansouri Rad, Karine Megerdoomian, and Rémi Zajac. 2000. Persian-English machine translation: An overview of the shiraz project. *Memoranda in Computer and Cognitive Science MCCS-00-319, NMSU, CRL.*

Majid Asgari-Bidhendi, Mehrdad Nasser, Behrooz Janfada, and Behrouz Minaei-Bidgoli. 2021. Perlex: A bilingual Persian-English gold dataset for relation extraction. *Scientific Programming*, 2021.

Yunes Azizian, Arsalan Golfam, and Aliye Korde Zafaranlu Kambuziya. 2015. A construction grammar account of left dislocation in persian. *Mediterranean Journal of Social Sciences*, 6(6 S2):98.

Mona Baker. 2011. *In Other Words: A Coursebook on Translation*. Taylor & Francis.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Guillaume Bonfante, Bruno Guillaume, and Guy Perrier. 2018. *Application of Graph Rewriting to Natural Language Processing*. John Wiley & Sons.

Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. 2006. Generating typed dependency parses from phrase structure parses. In *LREC*, volume 6, pages 449–454.

Marie-Catherine De Marneffe and Christopher D Manning. 2008. The Stanford typed dependencies representation. In *Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation*, pages 1–8.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. 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)*.

Pegah Faghiri and Pollet Samvelian. 2021. A corpus-based description of cleft constructions in Persian. *Faits de langues*, pages 183—206. Palancar, Enrique & Martine Vanhove (eds.).

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2021. ParsBERT: Transformer-based model for Persian language understanding. *Neural Processing Letters*, 53(6):3831–3847.

Abed Alhakim Freihat and Mourad Abbas. 2021. Proceedings of the second international workshop on nlp solutions for under resourced languages (nsurl 2021) co-located with icnlsp 2021. In *Proceedings of The Second International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2021) co-located with ICNLSP 2021.*

Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. When collaborative treebank curation meets graph grammars. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5293–5302, Marseille, France. European Language Resources Association.

Lifeng Han, Alan Smeaton, and Gareth Jones. 2021. Translation quality assessment: A brief survey on manual and automatic methods. In *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, pages 15–33, online. Association for Computational Linguistics.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486—2496.

S. Karimi. 2005. *A Minimalist Approach to Scrambling: Evidence from Persian*. Studies in generative grammar. Mouton de Gruyter.

Omid Kashefi. 2018. Mizan: A large Persian-English parallel corpus.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Mahsa Mohaghegh. 2012. *English-Persian phrase-based statistical machine translation: enhanced models, search and training: a thesis presented* *in fulfilment of the requirements for the degree of Doctor of Philosophy in Computer Engineering at Massey University, Albany (Auckland), New Zealand*. Ph.D. thesis, Massey University.

Behnoosh Namdarzadeh and Nicolas Ballier. 2022. The neural machine translation of dislocations. *ExLing 2022*, 28:127–131.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. 2018. Universal dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Mohammad Sadegh Rasooli, Manouchehr Kouhestani, and Amirsaeid Moloodi. 2013. Development of a Persian syntactic dependency treebank. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 306–314, Atlanta, Georgia. Association for Computational Linguistics.

Mohammad Sadegh Rasooli, Pegah Safari, Amirsaeid Moloodi, and Alireza Nourian. 2020. The Persian dependency treebank made universal. *arXiv preprint arXiv:2009.10205*.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Siamak Rezaei. 2000. *Linguistic and computational analysis of word order and scrambling in Persian*. Ph.D. thesis, University of Edinburgh. College of Science and Engineering.

John Richard Roberts, Behrooz Barjasteh Delforooz, and Carina Jahani. 2009. A Study of Persian Discourse Structure. Uppsala University Library.

John Robert Ross. 1967. *Constraints on Variables in Syntax*. Ph.D. thesis, MIT.

Mojgan Seraji. 2015. *Morphosyntactic corpora and tools for Persian*. Ph.D. thesis, Acta Universitatis Upsaliensis.

Mojgan Seraji, Filip Ginter, and Joakim Nivre. 2016. Universal dependencies for persian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2361–2365.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Jörg Tiedemann. 2016. OPUS – Parallel corpora for everyone. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products*, Riga, Latvia. Baltic Journal of Modern Computing.

Enric Vallduví and Elisabet Engdahl. 1996. The Linguistic realization of information packaging. *Linguistics*, 34(3):459–520.

Jan Wijffels. 2022. *udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit*. R package version 0.8.9.

S. Yousef and H. Torabi. 2021. *Intermediate Persian: A Grammar and Workbook*. Routledge Grammar Workbooks. Taylor & Francis.

Mina Zand Rahimi, Moein Madayenzadeh, and Mahdi Alizadeh. 2017. A comparative study of english-persian translation of neural google translation. *Iranian Journal of Applied Language Studies*, 9(Proceedings of the First International Conference on Language Focus):279–286.