Assessing Factoid Question-Answer Generation for Portuguese

João Ferreira
Centre for Informatics and Systems of the University of Coimbra, Portugal
jdcoelho@student.dei.uc.pt

Ricardo Rodrigues
Centre for Informatics and Systems of the University of Coimbra, Portugal
Polytechnic Institute of Coimbra, College of Higher Education of Coimbra, Portugal
rmanuel@dei.uc.pt

Hugo Gonçalo Oliveira
Centre for Informatics and Systems of the University of Coimbra,
Department of Informatics Engineering, Portugal
hroliv@dei.uc.pt

Abstract

We present work on the automatic generation of question-answer pairs in Portuguese, useful, for instance, for populating the knowledge-base of question-answering systems. This includes: *(i)* a new corpus of close to 600 factoid sentences, manually created from an existing corpus of questions and answers, used as our benchmark; *(ii)* two approaches for the automatic generation of question-answer pairs, which can be seen as baselines; *(iii)* results of those approaches in the corpus.

2012 ACM Subject Classification Computing methodologies → Natural language processing

Keywords and phrases Question-Answer Generation, Corpus, NLP, Portuguese

Digital Object Identifier 10.4230/OASIcs.SLATE.2020.16

Category Short Paper

Funding This work was supported by FCT's INCoDe 2030 initiative, in the scope of the demonstration project AIA, "Apoio Inteligente a Empreendedores (*chatbots*)".

1 Introduction

Natural language (NL) interfaces for computer systems are common these days. Many companies rely on *chatbots* for quick customer interactions, such as providing answers to common questions on a given domain, or providing links to documents where answers are likely to be found. Bearing this in mind, we present a new benchmark for question-answer generation (QAG) in Portuguese, followed by two baseline approaches. The corpus includes nearly 600 factoid questions, based on an existing question-answering (QA) corpus. The results of a QAG system are useful, for instance, for creating question-answer pairs, structured as frequently asked questions (FAQ) lists, which can be useful for QA agents or *chatbots*.

In the remaining document, we present a brief description of QA, question-generation (QG) and QAG, introduce the corpus, describe the used approaches and each of its modules, draw some conclusions, and reflect about future work.

9th Symposium on Languages, Applications and Technologies (SLATE 2020).
Editors: Alberto Simões, Pedro Rangel Henriques, and Ricardo Queirós; Article No. 16; pp. 16:1–16:9
OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

2 Background Concepts

Question answering is the process of automatically answering questions formulated in natural language, also using NL. It is a subfield of information retrieval (IR), but answers should contain precise information about the question, rather than a collection of documents. As in other subfields of IR, this often requires the semantic classification of entities, relations, or of semantically relevant phrases, sentences, or larger excerpts [13].

QA systems can use shallow or deep methods for analysing textual elements [1], and can be classified according to question-answer complexity, volume and quality of the source texts, type of corpora used, or how difficult it is to generate answers [12]. QA systems are also characterized by the type of user query – form, type and intent – and by the type of answers provided – named entities, phrases, factoids, links to documents, and summaries. QA can be further divided into open or closed domains, operating on structured data, such as databases or ontologies, or on unstructured data, such as text [12]. Although the latter is, arguably, more difficult, it is also the most common form of human produced documents.

Question generation [17, 2] deals with the analysis of text, identifying sentences and topics that are then used to formulate questions. QG is used in situations such as pedagogical environments, intelligent tutoring systems, conversation with virtual agents and IR [8].

Predictive questioning approaches to QA rely on the generation of questions from a raw corpus. They may assist humans in the creation of new questions, or in the selection of questions proposed by the system [7], possibly associated with the source excerpts. Foundations for such approaches lay on the creation of question-answer pairs, in the guise of a frequently asked questions system, suggesting multiple matches for a user question or presenting questions related to those, providing guidance in the process of selecting further questions on a given topic. QAG deals specifically with the generation of question-answer pairs [10], being used by QA systems or for assessing students, among others.

3 Data and Corpus Creation

Evaluating questions and answers is a complex process, mainly due to the fact that, for a single sentence, a large number of valid questions and respective answers can be generated. In order to do that, three elements are required: a target question, a target answer, and a factoid text that contains both of the previous. The latter is the information the system needs for generation, while the former two are the data required for evaluation. It is important to keep in mind that all these elements can usually be written in more than one way.

Multieight-04 [11] is a corpus of 700 factoid questions and respective answers, all available in seven languages, produced for the CLEF-2004 Multilingual QA Evaluation Exercise. It has two of the three elements required for assessing QAG: questions and answers. To add the third element, for each question-answer pair in Portuguese, we manually created a factoid sentence from which the pair could have been extracted. Examples are shown in Fig. 1.

No effort was made to balance how questions are created, but it is hard to say whether they would be balanced in real text. Moreover, requests not ending with a question mark ("?"), such as "*Mencione um bonecreiro.*", were disregarded. This resulted in a corpus of 581 entries, each including a factoid-like sentence, a question and an answer. This corpus can be used as a benchmark for assessing the automatic generation of questions and respective answers, in Portuguese. It is available at `https://github.com/jdportugal/multieight_pt_facts`.

Of course, multiple factoid sentences could have been produced for the same question. Likewise, a multitude of valid questions and answers may be produced for the same factoid sentence. By having a single possible variation of each element, the corpus will not be able

```
Factoid:  Umberto Bossi é o líder da Liga Norte.
Question:  Quem é Umberto Bossi?
Answer:  Líder da Liga Norte.

Factoid:  O prémio Nobel foi atribuído a Thomas Mann em 1929.
Question:  Em que ano foi atribuído o prémio Nobel a Thomas Mann?
Answer:  1929

Factoid:  Há 100.000 genes humanos.
Question:  Quantos genes humanos há?
Answer:  100.000
```

Figure 1 Examples of Multieight-04 question and answer, with added created factoid sentences.

to determine whether a generated question or answer that is different from the expected is indeed valid. Thus, assessment should be made on a minimum result basis; by that, we mean that the result obtained will be the worst score, with "real" results actually being better.

4 Question Generation Approaches

We compare two predictive approaches for QG in the created corpus: *(A)* exploits chunks and named entities, together with handcrafted rules (see Subsection 4.1); *(B)* relies on Semantic Role Labelling (SRL), using named entities in the process (see Subsection 4.2).

4.1 Approach A – Chunks, Entities, and Rules

Granularity-wise, syntactic chunks are a good option (if not the best) for breaking apart a sentence in blocks, easily rearranged for creating a question from an affirmative statement. This comes from the fact that, when transforming an affirmative sentence into an interrogative one, tokens in the same a chunk can usually be used as whole.

4.1.1 Chunks and Named Entities

This approach relies on the NLPPort suite of NLP tools [16], namely, for splitting sentences in chunks and for named entity recognition (NER). Both included models were trained in the Portuguese treebank Bosque 8.0 [6]. Resulting chunks are classified as nominal (NP), verbal (VP), prepositional (PP), adjectival (ADJP) or adverbial (ADVP) phrases. As for entities, they are classified as abstract, artprod (article or product), event, numeric, organization, person, place, thing, or time. Fig. 2 shows an example of a sentence and its chunks, then checked for the presence of named entities and replacing them by their type, later used for creating a question.

```
Factoid:  John L. Baird foi o inventor da televisão.

Chunks:  [NP John L. Baird][VP foi][NP o inventor][PP de][NP a televisão]
Entities (and types):  <START:person> John L. Baird <END>
→  [NP ?PERSON?][VP foi][NP o inventor][PP de][NP a televisão]
Rule (regex):  ^\[NP \?PERSON\?\].*

Question:  QUEM foi o inventor de a televisão?
Answer:  John L. Baird.
```

Figure 2 Example of the main constituents of Approach A.

As Fig. 2 also shows, generating the question is a matter of replacing the chunk with the entity, by, in that particular case, "QUEM" (WHO) – the interrogative pronoun for the type of the entity PERSON. The text of the answer is taken from the chunk that contains the entity. Of course, not all rules are this simple and not all the sentence structures are like this.

4.1.2 Rules

Rules are based on regular expressions to be applied on the chunked sentences, with entities replaced by their type, as in Fig. 2. When matching the rule, the pivotal chunk would also be the answer. Currently, 33 rules were written for addressing all the factoid sentences, including two generic rules that apply when no entity is found. In that case, the verb in the VP chunk is the main attribute for choosing which of the two rules to use: if the verb is "ser" (to be), and the VP chunk is surrounded by two NP chunks, the generated question is "O que é [left NP chunk] ?" (*What is ... ?*); if it is another verb, but also surrounded by NP chunks, the question is "O que é que [left NP chunk + verb] ?" (*What has ... ?*). In both cases, the answer would be the other NP chunk.

4.2 Approach B – Semantic Role Labeling and Named Entities

The second approach for QAG is based on SRL, the process of classifying words according to their semantic role in a sentence. Given its utility for this task, some approaches for QG already use SRL [4]. For this purpose, we relied on NLPNet [5], based on a Convolutional Neural Network (CNN) trained for performing SRL in Portuguese, and NLPyPort [3] for NER. The main idea was to exploit the *Argument* tags (Arg_0, Arg_1, . . . , Arg_n) and the *Verb* form in the sentence for generating the question, while the corresponding answer would be the argument or entity not used in the question. This aimed at creating simple, yet concise answers. Although, at first, this sounds like a feasible approach for QAG, sometimes, a single argument is found, or no arguments are found. Therefore, the proposed approach is divided in three strategies, all used in parallel for generating all possible questions and respective answers. A set of question-answer pairs is thus produced for each sentence, some better than others, i.e., some pairs are complete with a type in the beginning and a coherent wording, but others not so much.

4.2.1 Basic SRL Generation

The first strategy finds all arguments for the subject, all modifier arguments, and the verb, to then generate the question and the answer. QG follows the rule: `Verb + Argument + Modifier Arguments + "?"`, as depicted in Fig. 3, using the rule `Alternative Argument + "."` for generating the answer.

```
Factoid:  O prémio Nobel foi atribuído a Thomas Mann em 1929.

Arg_1:  O prémio Nobel                Arg_2:  a Thomas Mann
V: atribuído                          AM-Temp (Modifier Argument):  em 1929

Question:  Atribuído o prémio Nobel em 1929?
Answer:  A Thomas Mann.
```

Figure 3 Example of question generation using *basic* SRL (Approach B).

For the previous example, the alternative argument would be "A Thomas Mann", since it was not used in the question, and was thus considered to be the answer. Following this template, a question is generated for each of the arguments found, and the corresponding answer uses the arguments not included in the question. The result is an often incomplete question, because the first word, which usually characterises the subject of the answer, is missing and, without additional rules, is indeterminable.

4.2.2 Temporal and Spacial Generation using SRL

The second strategy is similar to the previous, but with a narrowed domain, by considering the sentences where a place or time argument is found. These arguments allow for determination of the type of subject in the generated answer – always a time/place argument – thus enabling the generation of complete questions. A rule for generating this type of question is `"Quando/Onde" + Verb + Argument 0 + Argument 1 + Modifiers Arguments + "?"`, as depicted in Fig. 4, using the rule `AM-TMP/AM-LOC + "."` for the answer. The generated question is close to the expected, but the verb complement "`foi`" is missing.

```
Factoid:  O prémio Nobel foi atribuído a Thomas Mann em 1929.

Arg_1:  O prémio Nobel                Arg_2:  a Thomas Mann
V: atribuído                          AM-TMP (Modifier Argument):  em 1929

Question:  Quando atribuído o prémio Nobel a Thomas Mann?
Answer:  Em 1929.
```

Figure 4 Example of temporal and spacial question generation using SRL (Approach B).

4.2.3 Entity and Rule Based Generation

When no arguments are identified by SRL, a fallback strategy is used. If an entity is found, its type is determined and the question is generated using a set of rules. If two entities are found and a rule using them exists, an "alternative type" is added to represent both entities' types, and the answer would be the entity not used in the question. One rule of this type for questions is `Type <verb> <Entity1>?` / `Alternative Type <verb> <Entity2>?`, with corresponding answers `Entity2` / `Entity`. An example is shown in Fig. 5.

```
Factoid:  Aleksandr Vassilievich Korjakov nasceu em Moscovo.

Entity:  Aleksandr Vassilievich Korjakov [PESSOA]      Entity:  Moscovo [LOCAL]
V: nasceu

Question:  Onde nasceu Aleksandr Vassilievich Korjakov?
Answer:  Moscovo.

Question:  Quem nasceu em Moscovo?
Answer:  Aleksandr Vassilievich Korjakov.
```

Figure 5 Question-answer pairs generated by the entity based fallback strategy (Approach B).

Since the number of rules is still limited, this part of the approach can be further improved, even if it is only useful when entities are present. Additional rules are shown in Fig. 6.

```
PESSOA LOCAL                              PESSOA PESSOA

Onde <verb> <Entity1>?                    Quem <verb> <Entity1>?
Quem <verb> <Entity2>?                    Quem <verb> <Entity1>?
```

Figure 6 Sample rules used in Approach B, on entity based back-up strategy.

5 Evaluation and Results

Systems based on both QAG approaches were assessed against the created corpus (Section 3). This assessment was based on the similarity between questions and answers generated for each factoid, and those in the corpus, given by two main metrics: BLEU [15], commonly used in the context of machine translation, but also for assessing generated questions [18]; and ROUGE, often used in the context of automatic summarization, for comparing the obtained summary with human-written summaries. Actually, we used ROUGE-L, a sub-metric of ROUGE that considers the longest matching subset [9]. BLEU relies on the comparison of n-grams in a candidate solution with n-grams of the expected solution, in a position-independent way. The more matches, the better the candidate solution is. Using only unigrams is equivalent to comparing the set of tokens used, which can lead to inflated results. Therefore, towards a complete evaluation, different values for n were considered, namely 1 to 4-grams, as well as a final average that considers the same weight for n-grams of each size.

We knew beforehand the limitations of this evaluation, the main of which is the existence of a single question-answer pair for a factoid, while many others were possible. However, manually creating a corpus with all the possible variations of each of those elements would be impractical. Rather, a single question-answer pair in the corpus was used, and all scoring was based on that pair alone. This outcome penalizes the system, because many valid questions and answers are rendered invalid. However, those that match the expected results are properly scored, and thus enable to assess the minimum performance of the system – not necessarily the best, because questions or answers that would otherwise be correct can be marked as incorrect. To better understand both the question and answer generation performance, results were computed for each of these elements, not for the pair as a whole.

In any case, if the system could not generate a question or an answer, it would be scored 0. Also, since both systems could generate more than a question-answer pair for a single factoid, two evaluation strategies were adopted: *(a)* considering the average of all the generated candidate solutions for each factoid, thus ensuring that over-generation of elements further apart from the expected is penalised; *(b)* in order to give a fair evaluation to the system, considering only the best generated question and ignoring the remaining.

5.1 Evaluation Figures

BLEU scores for QAG with each approach are in Table 1, and ROUGE scores are in Table 2. For any metric, Approach A, based on chunks, performs better than Approach B. This is not unexpected, because the first approach is more polished, due to the use of handcrafted rules, which are also in a larger number, allowing for the generation of more refined solutions.

Table 1 BLEU for the generation of questions and answers of both systems.

Evaluation	*BLEU n-grams*	Approach A	Approach B
average results for questions	1-gram	0.612	0.270
	2-grams	0.505	0.219
	3-grams	0.450	0.200
	4-grams	0.409	0.189
	weighted 1-4-grams	0.480	0.213
maximum results for questions	1-gram	0.614	0.358
	2-grams	0.507	0.297
	3-grams	0.451	0.273
	4-grams	0.411	0.258
	weighted 1-4-grams	0.482	0.289
average results for answers	1-gram	0.461	0.212
	2-grams	0.342	0.149
	3-grams	0.321	0.135
	4-grams	0.316	0.124
	weighted 1-4-grams	0.323	0.137
maximum results for answers	1-gram	0.463	0.279
	2-grams	0.344	0.201
	3-grams	0.323	0.181
	4-grams	0.318	0.165
	weighted 1-4-grams	0.324	0.185

Table 2 ROUGE for the generation of questions and answers of both systems.

Evaluation	***ROUGE-L***	**Approach A**	**Approach B**
average results for questions	Precision	0.438	0.164
	Recall	0.398	0.177
	F1	0.410	0.167
maximum results for questions	Precision	0.440	0.222
	Recall	0.400	0.239
	F1	0.412	0.225
average results for answers	Precision	0.345	0.210
	Recall	0.347	0.176
	F1	0.330	0.182
maximum results for answers	Precision	0.346	0.291
	Recall	0.348	0.240
	F1	0.332	0.291

5.2 Error Analysis

The main problems of Approach A are intrinsic: rules are handcrafted, even if they were thought to be as generic as possible. On average, a rule accounts for an excess of 15 generated questions (and corresponding answers) in the presented corpus. In the same corpus, there are 74 sentences with no questions generated for. Such sentences differ most from the other, in terms of structure, length and hence complexity – this can be a harbinger of the results to be expected in other types of text. Approach B fails more often. A common source of problems is in the SRL system used, NLPNet: some semantic roles are incorrectly classified and result in incorrect solutions; other roles are simply not classified, forcing the system to use the fallback strategies, which are more prone to errors and often generate results of lower quality. An example of a SRL-related problem is depicted in Fig. 7, where bad question and answer were the result of mislabeling the `Arg_0` and `AM-LOC` arguments.

```
Factoid:  A melhor maneira de combater as alergias é administrar, em quantidades
mínimas, as substâncias que causariam ao paciente reacções alérgicas.

Arg_0:  que                             Arg_1:  ao paciente reacções alérgicas.
AM-LOC: em quantidades mínimas.         V: causariam

Question:  Onde causariam que ao paciente reacções alérgicas?
Answer:  Em quantidades mínimas.
```

Figure 7 Sample questions and answers generated by the entity based fallback strategy.

Even the fallback strategies can fail to generate a question, thus resulting in a penalisation when computing the score. For instance, no questions (nor answers) are generated for the factoid "*O aumento da população mundial por ano é de 94 milhões.*", because no semantic roles were assigned and no useful entity types were found. The lack of rules was one of the main problems of the fallback entity based strategy.

To overcome the noted flaws, we could start by improving SRL, so that less solutions depended on the fallback strategies. NLPNet dates back from 2013 and relies on a CNN. Yet, since then, more powerful architectures of neural network were developed and used on sequence-labelling tasks. Those include Recurrent Neural Networks with LSTM layers or Transformers [14]. Handcrafting more rules for the entity based solution generator of Approach B would also allow for the generation of better solutions.

Overall, we believe that Approach B has more room for improvement. Though, it may be a good idea to develop an hybrid system that combines the best parts of the two approaches.

6 Conclusion

We have presented a corpus with 581 factoid sentences, in Portuguese, each with a corresponding question-answer pair. The corpus was used for testing two approaches for QAG, applicable, for instance, to the automatic creation of FAQ-style lists from raw documents, and for automatic population of websites or knowledge bases for NL interfaces, including *chatbots*. Both approaches are rule-based, with rules operating on the output of: *(i)* a syntactic chunker and named entity recogniser – with the best results; and *(ii)* a semantic role labeller. Even though the corpus is not that complex, as most sentences use a straight subject-verb-object structure, we do feel that, as we have done for comparing our approaches, it can be used as a benchmark for other researchers working on QAG for Portuguese.

In the future, it is our intention to apply the presented approaches, or improved versions, to more complex text. In fact, we have already performed preliminary experiments on law text, but results were unsatisfying. Yet, we see the work presented here as a good starting point, where we can build on for pursuing our goal. This includes not only improving the proposed approaches, but also testing more recent machine learning approaches, which would avoid the manual creation of rules, a time-consuming process.

References

1 Johan Bos and Katja Markert. Combining Shallow and Deep NLP Methods for Recognizing Textual Entailment. In *Proceedings of the Pascal Challenges Workshop on Recognising Textual Entailment*, Southhampton, UK, April 2005.

2 Daniel Diéguez, Ricardo Rodrigues, and Paulo Gomes. Using CBR for Portuguese Question Generation. In *Proceedings of the 15th Portuguese Conference on Artificial Intelligence (EPIA 2011)*, pages 328–341, Lisbon, Portugal, October 2011. APPIA.

3 João Ferreira, Hugo Gonçalo Oliveira, and Ricardo Rodrigues. Improving NLTK for Processing Portuguese. In Ricardo Rodrigues, Jan Janoušek, Luís Ferreira, Luísa Coheur, Fernando Batista, and Hugo Gonçalo Oliveira, editors, *Proceedings of 8th Symposium on Languages, Applications and Technologies (SLATE'19)*, OpenAccess Series in Informatics, pages 18:1–18:9. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, June 2019.

4 Michael Flor and Brian Riordan. A Semantic Role-Based Approach to Open-Domain Automatic Question Generation. In *Proceedings of the 13th Workshop on Innovative use of NLP for Building Educational Applications*, pages 254–263, 2018.

5 Erick Fonseca and João Luís Rosa. A Two-Step Convolutional Neural Network Approach for Semantic Role Labeling. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2013.

6 Cláudia Freitas, Paulo Rocha, and Eckhard Bick. Floresta Sintá(c)tica: Bigger, Thicker and Easier. In *Proceedings of the 8th International Conference on Computational Processing of the Portuguese Language (PROPOR '08)*, pages 216–219. Springer-Verlag, 2008.

7 Sanda Harabagiu, Andrew Hickl, John Lehmann, and Dan Moldovan. Experiments with Interactive Question-Answering. In *Proceedings of the 3rd Annual Meeting of the ACL (ACL '05)*, pages 205–214, Morristown, New Jersey, USA, 2005. ACL.

8 Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. A Systematic Review of Automatic Question Generation for Educational Purposes. *International Journal of Artificial Intelligence in Education*, 30:121–204, 2020.

9 Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of Workshop on Text Summarization Branches Out, Post2Conference Workshop of ACL*, 2004.

10 Bang Liu, Haojie Wei, Di Niu, Haolan Chen, and Yancheng He. Asking Questions the Human Way: Scalable Question-Answer Generation from Text Corpus. In *Proceedings of The Web Conference 2020 (WWW '20)*, pages 2032–2043. IW3C2, 2020.

11 Bernardo Magnini, Alessandro Vallin, Christelle Ayache, Gregor Erbach, Anselmo Peñas, Maarten de Rijke, Paulo Rocha, Kiril Ivanov Simov, and Richard F. E. Sutcliffe. Overview of the CLEF 2004 Multilingual Question Answering Track. In *Multilingual Information Access for Text, Speech and Images, 5th Workshop of the Cross-Language Evaluation Forum (CLEF), Revised Selected Papers*, volume 3491 of *LNCS*, pages 371–391. Springer, 2004.

12 Mark T. Maybury, editor. *New Directions in Question Answering*. AAAI Press and The MIT Press, Menlo Park, California, and Cambridge, Massachusetts, USA, 2004.

13 Marie-Francine Moens. *Information Extraction: Algorithms and Prospects in a Retrieval Context*. Springer-Verlag, Berlin Heidelberg, 2006.

14 Hiroki Ouchi, Hiroyuki Shindo, and Yuji Matsumoto. A Span Selection Model for Semantic Role Labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1630–1642. ACL, 2018.

15 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 on ACL*, pages 311–318. ACL, 2002.

16 Ricardo Rodrigues, Hugo Gonçalo Oliveira, and Paulo Gomes. NLPPort: A Pipeline for Portuguese NLP. In *Proceedings of the 7th Symposium on Languages, Applications and Technologies (SLATE'18)*, OpenAccess Series in Informatics, pages 18:1–18:9, Germany, June 2018. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing.

17 Vasile Rus and Arthur C. Graesser. The Question Generation Shared Task and Evaluation Challenge. Workshop Report, The University of Memphis, 2009.

18 Xingdi Yuan, Tong Wang, Caglar Gulcehre, Alessandro Sordoni, Philip Bachman, Sandeep Subramanian, Saizheng Zhang, and Adam Trischler. Machine Comprehension by Text-to-Text Neural Question Generation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 15–25, Vancouver, Canada, 2017. ACL.