School of Informatics

Informatics Research Review
Unsupervised Learning on Corpus-Based GI

[redacted]

January 2021

Abstract

Grammar induction, as one of important areas in natural language processing, has drawn attentions for years. Unsupervised grammar induction is now a more attractive field in that area, compared with supervised grammar induction which is limited by the availability of training data. This paper concludes papers at the frontiers of unsupervised grammar induction fields and will detailed introduce the motivation of unsupervised grammar induction and several impressive learning models.

Date: Friday 22nd January, 2021

Supervisor: [redacted]

1 Introduction

The whole section is divided into two subsections through thematic approach. The first subsection includes the description of the chosen topic, motivations as well as the issues this paper wants to solve, and the organisation of this paper. In the second subsection, there are brief introductions about the main related literature used in this paper and a paragraph briefly stating the general findings to the issues as well as some similarities among the three paper.

1.1 Topic description & Organisation

The topic of this paper, Unsupervised learning on corpus-based grammar induction, is not a theory nor an argument but a scope of related literature to be investigated. It mainly involves two fields, one is machine learning and the other is linguistics. Its explanation is to generate a language model from a given corpus, based on a specific machine learning method, unsupervised learning. Through the model, it is possible to automatically analyze the grammar of given test texts, including but not limited to Part-of-Speech (POS) tags of words.

This paper intends to introduce unsupervised methods which are used in grammar induction (GI) task, thus, it focuses on two issues. One is the motivation and advantages of developing these unsupervised learning methods on GI, which will be explained in the Sec 1.2 and Sec 2. The other is specific details of these methods on how these algorithms generate required language models, which will be listed and detailed explained in Sec 2.

The following is the outline of this paper. Sec 1 is an introduction of the whole paper including summaries of the three mainly cited paper, Sec 1.2. While Sec 2 will be more professional and include more specific methods, such as dependency structure of grammar in [1] and genetic-based inside-outside algorithm to train language model in [2]. Sec 3 integrates all the findings from literature into a brief paragraph and generates a proper conclusion which answers the issues came up in the above paragraph.

1.2 Related surveys

Considering GI is a popular topics, there are plenty research papers related to this field. However, most of them are based on supervised learning[3]. In the past decades, there has been some research papers trying to use unsupervised learning methods to solve GI problem. From all research paper, it can be found that the motivation of using unsupervised methods is account to some linguistic issues, such as the population of the language and the theoretical limitation of grammar[3].

In this paper, apart from trivial arguments from other documents, there are mainly three articles cited as evidences. Evidences include the motivation of unsupervised learning generated by the author and some improved unsupervised learning models. And due to the issuance times are different and the topic of papers is highly relevant, there is no surprise to find some papers' descriptions of motivation and part of the specific experimental details are inherited from others. Following are brief summaries of each paper, more details of the methods mentioned will be shown in Sec 2.

1. [1] presents an improved generative model, based on combining two existing methods, for the unsupervised learning of dependency structures. It also evaluate the model with other baseline methods on multi-languages to show the features extracted by the model

is convincing and is robust cross-linguistically. Apart from the above method, this paper also provides a comprehensive overview of the development of unsupervised learning in natural language processing.

2. [3] mainly describes a new method based on the automatic extraction of certain patterns, words, in the texts. Therefore, it is possible to identify constituents in the sentence through the classes of the extracted words. What needs to be emphasized is this paper only provides a model to extract words without the continuous steps mentioned above, and the training data is also labelled but only annotated with POS tags.

3. [4] offers a new point of view to implement GI. In stead of analyzing monolingual corpus like other methods, it analyzes bilingual parallel corpora, and with the help of comparisons of the two language to make the model achieve better analytical abilities. In order to make the alignment of the two languages 'visible', that is, appropriate to add it to the model generated through baseline methods, this paper comes up with a Bayesian model to describe the alignment. Moreover, This paper also involves the discussion about the advantages and disadvantages of the above integrated method.

2 Literature Review

The knowledge from relevant literature forms three parts according to thematic approach. In the first subsection Sec 2.1, there are brief introductions about the relevant fields' background. The second subsection, Sec 2.3, will conclude the motivation from the necessity and benefits on using unsupervised method. The last subsection will list some detailed unsupervised learning techniques and do some comparisons to make the requirements of those techniques as well as pros and cons clear. Although the three main paper mentioned in Sec 1.2 can cover the main ideas of each subsection, other papers only relevant to part of the topic are also cited to make this paper comprehensive and convincing.

2.1 Backgrounds of relative fields

In machine learning, there are mainly three types of learning method: supervised learning, unsupervised learning and reinforcement learning. For supervised learning, data is usually annotated, such that the machine knows what to look at when training models. In [5], supervised learning is likened as a hunting dog knowing the smell of targets. For instance, in a task of recognising traffic lights from pictures, traffic lights are circled out in the training data to help machines to learn the feature[1], thus, machines can apply the feature to other pictures to confirm if there are traffic lights. On the contrary, in unsupervised learning, the training data is given without any label, therefore, from training data the machine will try to get whatever features that seem to be relevant to make decisions[5]. That is, in this type of learning, features will not be directly given ahead but the machine should find the features, which might be the traffic lights' color and the shape, itself instead. Reinforcement learning, the latest frontier of machine learning, is based on incentive mechanisms and makes the machine to do numbers of trials to get final objectives. The incentive mechanisms reward the machine when it approach the objective, otherwise punish it[5]. Instance of reinforcement learning includes AlphaGo, path guiding systems in video games and etc.

[1]Here the feature means traffic light

Natural language processing (NLP) is a way to quickly select required knowledge and useful information from texts on Internet and other places. As [6] presents, it is becoming increasingly difficult for human to discover needed knowledge/wisdom from the numerous natural language texts in the connected world. By now, NLP has become one of the most popular fields in machine learning and there are many NLP applications such as translators, AI assistant[2] and etc. It is worth mentioning that most NLP applications are closely related to syntax and grammar. Therefore, there is no surprise to draw the conclusion that GI, learning syntactic structure from given texts, is one of core tasks in NLP[6]. Due to the flexibility of natural language, it is difficult to define a grammar structure for most kinds of languages. From [7], it is recognised as possible that some methods designed for a certain field are useful in other fields. Context-free grammars used to be a standard tool for defining syntax of programming language[7]. However, now in natural language, it has become an important tool to identify particular grammar structures as context free grammar (CFG) structure and widely used in many GI tasks.

2.2 Reinforcement and Supervised learning approaches in NLP

Reinforcement learning, as a latest frontier of machine learning, has already been used in a few papers to solve NLP problems in past two years. Google's BERT[3] system is developed through reinforcement learning and it is the most recent breakthrough[4] in NLP field as it is the first NLP system that beats human performance[5] in both two measurement, EM and F1[6], according to the SQuAD1.1[7] test[8]. From [9], it is shown that BERT is still based on unsupervised methods and the main difference is that BERT use network structures to do a great number of trials instead of using traditional generative model. As a emerging technology, reinforcement learning still needs more time to be mature and is one of the directions worth exploring in the future[10]. In this paper, both methods in [9] and [10] will be briefly introduced in Sec 2.4 to compare with traditional unsupervised learning method.

Different from reinforcement learning, because of the mature theory and technology, most of GI problems are handled through supervised learning[3] and there are more approaches that can boost computational efficiency in supervised learning to select than any other two machine learning methods[2]. Both [2] and [3] mention that supervised approaches usually work on a annotated corpora, or treebank as a reference to extract the grammar. The annotated dataset or treebank is essential because only with such dataset, machines can directly capture the features in the training dataset, for instance the dependency structure between POS tags[1] and probabilities of words' POS tags[3], and apply them into test dataset. However, the annotated data is also the constrains of supervised learning approach, because machine learning requires large amount of labelled data to collect adequate features while the size of annotated dataset or treebank is limited due to man power issues[2]. There are some attempts to deal such problems, for instance, [11] comes up with a solution to use less features in supervised learning to decrease the cost on building the large corpora, [12] provides a solution to use bitext[8] with a source language parser to constrain the target model's 'size' or space and therefore decrease the size of required annotated dataset.

[2]This assistant is used to answer users' questions, one of the instances is Siri
[3]Bidirectional Encoder Representations from Transformers
[4]Now BERT system is the top20 in SQuAD2.0 and Top5 in SQuAD1.1 leaderboard.
[5]This performance is made by Stanford University.
[6]EM is Exact Match, F1 is F1 score
[7]The Stanford Question Answering Dataset.
[8]Bilingual texts with word level alignment.

2.3 Motivations of unsupervised GI

There are many necessities of using unsupervised learning to handle GI problems due to limitations of the above two types of machine learning:

1. From the above paragraphs, it is clear to notice, some NLP reinforcement learning approaches nowadays are still inspired by the unsupervised learning theory. Therefore, paying more attention to the unsupervised learning approaches is laying the foundation for the approaches of reinforcement learning t some extent.

2. As for supervised learning mentioned above, although there are some solutions to reduce the need of annotated dataset, the problem of the high cost of annotating dataset remains unresolved[2].

3. Apart from the high cost, only some languages such as English, Spanish, Chinese and etc, own sufficient number of treebank, while many other languages may not have treebank available[3]. Those languages are minority languages and ancient languages such as Persian and usually not only lacks the annotated sentences but also lacks the people to annotate the data[2].

4. According to [3], another limitation of treebank is that treebank is usually composed of some particular types of content, such as Wikipedia articles and newspaper articles. Therefore, the final trained model nay not be appropriate to other types of contents, for instance novels, poems and articles in particular areas.

5. Besides, there is a minor theoretical limitations for GI as the fact that grammars sometimes cannot be correctly identified from positive examples alone[13]. However, under normal circumstances of supervised learning, this point is not taken into account.

To conclude, the main reason that unsupervised GI is getting popular is that under normal circumstances, it does not require a treebank which has many limitations to train the language model. However, apart from this necessity, using unsupervised learning can benefit in many aspects and there are many research papers focusing on unsupervised learning with various reasons[1]:

1. In [14], it proves the poverty of the traditional unsupervised learning algorithms with sufficient empirical statistics. This provides the direction for the follow-up researchers to improve the performance of the unsupervised learning approaches at that time.

2. Unsupervised learning can help to build a large treebank as a first stage[1] or increase the size of treebank in some extent. In [15], unsupervised learning is used to exact similar constituent[9] in multiple sentences[10] and try to replace the constituent with each other in order to get new sentences. In this way, the size of treebank is enlarged.

3. There are more than one models can be used in NLP. Supervised learning approaches focus more on dependency structure models while there are another models, such as HMM[11]

[9]Constituents are language units, such as phrases like "walk a dog", a phrase like "a dog" or a word like "dog", in which we can arrange the structure of a sentence.

[10]Here, we still need treebank as input, but we do not require its labels as we are using unsupervised learning methods.

[11]Hidden Markov Model

and Bayesian Model, which are more suitable and have higher performance through unsupervised learning. As [1] recommends, there are papers using different models with unsupervised learning aiming at achieving higher model performance.
[16] uses hidden Markov process through unsupervised learning to analysis speech data and constructs automata to better deal with ambiguities in speech recognition.
[17] uses greedy search algorithm in Bayesian framework to train a language model which beats n-grams models[12].

4. Objectives and approaches in unsupervised learning are various, but in [1], it divides them into two types. One is working on 'weak' generative capacity of models, such as [16] and [17]. Their models' hierarchical structures are not useful until models are utilized to improve the performance over some traditional observed structures. On the contrary, the other type is working on 'strong' generative capacity of models, such as [14], [15] and [1]. They continuously focus on the models' structures from introducing the model to the improvement the model. Therefore, the former two papers are more theoretical and pay more attention on the basic logic of the model such as introducing the hidden Markov process and Bayesian framework, while the later three papers are more practical and easier to reproduce the models.

2.4 Unsupervised learning techniques in NLP

Here will introduce some unsupervised learning approaches in detail, including models' theory and requirement of using. Meanwhile, one of reinforcement learning NLP systems, BERT, will also be briefly introduced and compared with models generated through unsupervised approaches. The empirical steps as well as results of those models will not be displayed. There is no reason to compare such details with each other because each paper uses different evaluation criteria and the objective of each approach is not the same. Moreover, it is worth mentioning that, these subsection is based on [1], [3] and [4], but some other papers will be added for comprehensive and convincing purposes.

1. The first model comes up in [1], and is called unsupervised dependency parsing. Dependency parsing, as a common structure of NLP application, has been widely used in both supervised and unsupervised learning. Most of unsupervised parsing remains tree or phrase structure grammar, such as the generative distributional model in [18]. Tree or phrase structure means the sentence will be divided in to several constituents recursively until each constituent only includes a single word. This kind structure is hierarchical and therefore suitable for showing dependency relationships[1]. However, in the new unsupervised parsing provided in [1], a new structure, function-argument and modification structures, replaces the tree structure due to three reasons[1]:

   (a) Supervised parsing not only applies word-class level information but also make use of lexical information which can also be taken into consider when building unsupervised parser.

   (b) Directly induce the function-argument and modification structures will be advantageous as what tree structures want to present is such dependencies.

   (c) Some languages, such as Chinese, do not contain much functional words, making the word-class information less clear, therefore, tree structure model generated only through word-class level information would be inaccurate.

[12] N-gram model is one of traditional supervised models.

In the traditional tree structured dependency parsing, it is incapable of encoding even first-order valence facts. As the example given in [1], if given "NOUN NOUN VERB" sequence to models, both nouns will be attached to the verb[13], because all traditional dependency models cannot realize the fact that there can only be one subject. To fix this limitation, [1] come up with an improved dependency model by picking up the key idea, head outward process[11], in supervised learning field. Moreover, [1] also construct a model of valence into the improved model which is used for checking whether the generated dependency structure obeys the rules in the valence model when applying head outward process from the ROOT of the generated model structure[11]. [1] has proved this model works successfully and better than other traditional unsupervised learning models. It is also confirmed that the improved dependency model works cross-linguistically and can be integrated with another state-of-art, distributional constituency induction model, by summing over lexical information from these two models to achieve a even higher accuracy on showing dependency relationships.

This improved model is able to minimize the number of hidden structure that must be induced, therefore, only required a modest amount of training data, the dependency structure can be recovered successfully[1]. However, the improved model is still somewhat problematic, because it forces the constituents' structure, which should be arbitrary, to be x-bar-like. Otherwise, the identified dependency relationships would be somewhat conflicting with each other and therefore cause inaccuracy in valence model[14].

2. [3] makes an assumption that in a sentence, there exists some classes of words marking the beginning or the end of constituents called separators or sub-separators, such as Modal, Preposition and etc. Separators, such as Modal, are words or classes have great influence in determining the sentence structure and we can break the whole sentence into constituents at these words' positions to generate a new level in a parse tree. For instance, there is a sentence "My brother can catch a bird", "can" is a separator as it can divide the whole sentence into "My brother" and "catch a bird" two new constituents, which means the parse tree of this sentence generate a new level[15]. Sub-separators, on the other hand, can only indicate the beginning or the end of the constituent it belongs to and if break the sentence as above, there will not be new level generated in a parse tree. For instance, "a" in the above example is a sub-separator, because "a" is the beginning of "a bird" but when breaking the constituent from it, it cannot lead to a new level in the parse tree [16]. Because this model is relevant to parse tree, this generative model remains in tree structure.

   In supervise learning, we need to manually find the separators from annotated treebank, but in unsupervised learning which is [3] does, we hope the machine can find the separators itself from the annotated treebank[17]. This model directly works on the treebank and the input of the model are sequences of POS tags instead of real sentences. Based on those POS tags and relevant papers, this paper makes following sensible assumptions which are

[13] If one constituent is relevant to another, than we can attach them through an arc to show there dependency relationship. Here, NOUN, NOUN, VERB are all words which can also be treated as constituents.

[14] one of the parts of the improved model

[15] "My brother" is a NP (Noun Phrase) and "catch...bird" is a VP (Verb Phrase), both are non-terminals. Non-terminal means this constituent can be divided into further level.

[16] "a bird" is a NP (Noun Phrase) but after being broken, it will become a terminal "bird" which cannot be further divided in the parse tree, meaning the parse tree will not grow a new level.

[17] To mention that, here annotated means the treebank has already be marked POS tags. Considering the features of this learning is separators, the marked POS tags cannot be seen as labels, and this learning type is still unsupervised learning.

the key ideas of this model[3]:

(a) Assume the most frequent sequence in the treebank is a constituent, called safe constituent (SC). This assumption is made through experience as under normal circumstances, the occurrence number of a frequent constituent is much higher than an arbitrary sequence.

(b) Based on the first assumption and [19], this paper assume any sequence that has the same beginning, L_{sc}, and ending, R_{sc}, as SC is a constituent. [19] presents a fact that longer constituents usually contain short, common equivalents. Constituencies of those equivalents can be easily found as a grammar rule. For instance, in two long sentences, it is usually correct to make the judgement that both sentences there are "NP" and "VP" acts as subject and predicate[18].

(c) Based on the above assumptions, [3] compares the number of co-occurrences of L_{sc} and E to the number of co-occurrences of R_{sc} and E to determine which tag, L_{sc} or R_{sc}, appears more important to E. Then, [3] makes another assumption on identifying whether E is separator or sub-separator based on the position of E corresponding to the tag chosen above and their co-occurrence.

This model can only generate POS tags which are separators, but cannot directly recognize the constituents of a sentence compared with the first model. In fact, as long as the separators are confirmed, identify the constituents can be easily done by taking the classes of words between separators. Another limitation is this model still needs treebank as training data, that is, although it does not need to spend man power on labelling the data, it is still limited by the languages and contents of the treebank. For instance, all the data used for training is from Wall Street Journal, and this model has not been tested on whether it can work cross-linguistically as the first model. However, this model does offer a whole new perspective on unsupervised GI, that is instead of doing a step-by-step analysis on words in the sentences, looking for patterns in the distribution of representative words' classes[19]. Just as the [3] said in the conclusion, in analysing POS tags' regulations of distribution, there are more work need to be done, such as taking into the occurrences of each POS tag into consideration.

3. Unlike the above two models working on monolingual language at each time, the model in [4] requires a bilingual parallel text[20] as input and [4] introduces **unsupervised** bilingual cues[21] to improve the model performance. This idea is inspired by other linguistic induction tasks, such as lexicon acquisition[20], morphological segmentation[21] and POS tagging[22]. The premise of this approach is that, some ambiguities in one language may not happen in another language. For instance, in English sentences, there are usually ambiguities around the object of a prepositional phrases, such as "I saw [a man [with a telescope]]" and "I saw [a man] [with a telescope]". In other languages, Chinese for instance, the above sentence will be translated as "I saw [a [with a telescope] man]" for the first meaning and "I [with a telescope] saw [a man]" for the second meaning. It can be seen that the ambiguity which originally existed in English has disappeared in Chinese. Therefore, through parallel bilingual text, some grammar ambiguities can be avoided from being misidentified by the machine.

[18] In fact, "VP" can not only include predicate such as "catch a bird", but include a predicative such as "is fat".

[19] Such as POS tags.

[20] Parallel here means the texts in two languages are aligned in word level like the correspondence of translation.

[21] Like the second model above, only part of the model is using unsupervised technique.

As the above example exploits, although the ambiguity can be avoided when comparing Chinese with English text, orders of the words and phrases in Chinese and English are different. Besides, there are other challenges when considering how to discover cross-lingual patterns in data[4]. Thus, [4] adapts a formalism called unordered tree alignment[23]. In [23], the alignment tree can allow arbitrary parts in any two trees to diverge in structure with the guarantee that their grammatical structure being preserved. Another attractive advantage of alignment tree is the linear computational time. After the alignment tree for the two languages being constructed, there is a Bayesian generative model used for generating constituents for each language and combining them to pairs of bilingual paralleled constituents to check the ambiguities and other issues. Then, we can get parallel sentences assembled from those constituents and words-level alignment as final results[4].

Compared with first and second statistical models, the use of bilingual text makes it works more cross-linguistically. Moreover, because of the words-level alignment, [4] can avoid some kinds ambiguity, which can be hardly achieved by monolingual models. However, this methods will be constrained by the training dataset, including the size, contents, and language choices of the bilingual parallel text.

4. Different from all three kinds of models above, frankly speaking, BERT is a reinforcement learning model. However, according to the [6], this model is based on unsupervised learning theory somehow and integrates some supervised learning data as well. There are several latest technologies implemented in BERT that distinguish it from other unsupervised learning models.

   (a) BERT implements unsupervised feature based approaches. Pre-trained word embedding, as an important part of modern NLP systems, can greatly improve the performance than embedding from scratch[24]. Usually, these word embedding vectors are pre-trained through left-to-right language modeling objectives[25]. Just as the [6] shows, BERT is designed to pre-train deep bi-directional representations from **unlabelled dataset** by considering both left and right context in each layer.

   (b) Another method used in BERT is unsupervised fine-tuning approaches. In BERT, most parameters can be pre-trained ahead, and it is possible to generate models which are suitable for a wide range of tasks, such as translation and Q&A interaction, through an additional changeable output layer to do fine-tuning[6]. The advantage of this architecture is that only few parameters need to be learned from scratch[26]. Therefore, the re-usability of BERT is much higher than the three models above.

   (c) Transfer learning from supervised data also contributes to the training of BERT system. This transfer learning is implemented to make BERT more effective as according to [27, 28], natural language inference and machine translation through supervised learning in large dataset help to transfer to BERT system.

   BERT model's architecture is a multi-layer bidirectional Transformer encoder[6]. This network is based on the description in [29] and compared with the original transformer that can only encode word from left to right, the transformer in BERT is bi-directional. To make bi-directional representation available, a 'MASK' is used to randomly hide words in the text and make the network to guess the content under the 'MASK'[6]. The above are built in layers ahead of the output layer, and works in pre-training step which usually takes a lot of time. The output layer is responsible for fine-tuning, where take the control of what types of models will be generated. At this stage, we need to control the input and output according to specific model types[6]. For instance, paraphrasing task requires sentence

pairs as input and output; Q&A task requires question-passage pairs etc. Moreover, the adjustment of hyper-parameter is included in fine-tuning tasks.

This reinforcement learning has gained a lot of praise, but there are still some drawbacks. One of them is the 'MASK' tag using in the model will lead the inconsistency between pre-training and fine-tuning, stated in [6]. Although it states that only 1.5% of words happen to be arbitrary replaced in the text which will not affect the models' ability of understanding the grammar, the model may become more accurate if this is solved. Another drawback is the model will need more time to converge during pre-training as some of the words are replaced by 'MASK' and compared with the left-to-right transformer, BERT will be slower[6].

3 Summary & Conclusion

With new learning technologies emerging one after another, there are more and more approaches targeting on GI (grammar induction) task. Thanks to manual annotating the dataset, supervised learning methods used to be more accurate and therefore very popular at NLP early stage. Until now, some supervised methods and tasks are embedded into semi-supervised, unsupervised and reinforcement methods, such as the transfer learning in BERT[6], which can improve efficiency. However, with the potential limitations of language itself and annotating dataset[3, 2], there are more and more researches working on unsupervised approaches. At present, reinforcement learning, as a emerging technology, relies more on unsupervised learning theories and supervised learning experience in NLP field. For instance, BERT embeds two unsupervised approaches and refers many supervised tasks for transfer learning[6].

Supervised GI is mainly constrained to the treebank. The limitations of treebank include the lack of language choices, the lack of content types and the constrain on treebank size. Unsupervised methods on the other hand owns unlimited training data in theory[6]. The motivation of unsupervised GI also goes to the benefits that unsupervised GI can bring, such as helping to enlarge the size of treebank, improving some specific language models and etc.

With the passion of researching unsupervised GI, there are numerous relevant research papers where three of them are chosen to have a deep look at models they design. They are quite various in aims, theory and structure. The first one basically improves from a traditional dependency model and through combining it with another existing model to achieve higher performance. The second model is from a new perspective, analysis of POS tags' regulation in the given data, to generate sentences. The third model takes the language differences into consideration and achieves disambiguation by looking at words-level alignment in bilingual parallel text. The forth model, BERT, is in multi-layer network structure to do pre-training and fine-tuning tasks through unsupervised approaches.

The forth model achieves the highest score from SQuAD1.1 among all models[22] and it is the first NLP model that beats the human performance with a narrow lead (less than 1.0%). By now, BERT only rates 7_{th} on SQuAD1.1 on the leaderboard and the 1_{st} model reaches 90% in EM and 95% in F1 score, both exceed human performance over 4%. There is no doubt that there will be more models generated through reinforcement learning methods and the performance will be more accurate.

[22]BERT broke the record in 2019 on SQuAD1.1, while other models appeared earlier than that.

References

[1] Dan Klein and Christopher D. Manning. Corpus-based induction of syntactic structure. pages 478–es, 2004.

[2] Mohsen Arabsorkhi, Hesham Faili, and Mansoor Zolghadri Jahroumi. Using genetic algorithm for Persian grammar induction. *2009 International Conference on Natural Language Processing and Knowledge Engineering, NLP-KE 2009*, 2009.

[3] Jesús Santamaría and Lourdes Araujo. Identifying patterns for unsupervised grammar induction. *CoNLL 2010 - Fourteenth Conference on Computational Natural Language Learning, Proceedings of the Conference*, (July):38–45, 2010.

[4] Benjamin Snyder, Tahira Naseem, and Regina Barzilay. Unsupervised multilingual grammar induction. *ACL-IJCNLP 2009 - Joint Conf. of the 47th Annual Meeting of the Association for Computational Linguistics and 4th Int. Joint Conf. on Natural Language Processing of the AFNLP, Proceedings of the Conf.*, (August):73–81, 2009.

[5] Karen Hao. What is machine learning? `https://www.technologyreview.com/2018/11/17/103781/what-is-machine-learning-we-drew-you-another-flowchart/`. MIT Technology Review. Online; accessed 16 January, 2021.

[6] KR Chowdhary. Natural language processing. In *Fundamentals of Artificial Intelligence*, pages 603–649. Springer, 2020.

[7] Jukka Paakki. Attribute Grammar Paradigms—A High-Level Methodology in Language Implementation. *ACM Computing Surveys (CSUR)*, 27(2):196–255, 1995.

[8] Stanfoord NLP Group. Squad2.0 the stanford question answering dataset. `https://rajpurkar.github.io/SQuAD-explorer/`. Stanfoord NLP Group. Online; accessed 16 January, 2021.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[10] Tianyang Zhang, Minlie Huang, and Li Zhao. Learning structured representation for text classification via reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[11] Rebecca Hwa. Supervised grammar induction using training data with limited constituent information. *arXiv preprint cs/9905001*, 1999.

[12] Kuzman Ganchev, Jennifer Gillenwater, and Ben Taskar. Dependency grammar induction via bitext projection constraints. *Lab Papers (GRASP)*, page 43, 2009.

[13] E Mark Gold. Language identification in the limit. *Information and control*, 10(5):447–474, 1967.

[14] Alexander Clark. Unsupervised induction of stochastic context-free grammars using distributional clustering. In *Proceedings of the ACL 2001 Workshop on Computational Natural Language Learning (ConLL)*, 2001.

[15] Menno Van Zaanen. Abl: Alignment-based learning. *arXiv preprint cs/0104006*, 2001.

[16] James K Baker. Trainable grammars for speech recognition. *The Journal of the Acoustical Society of America*, 65(S1):S132–S132, 1979.

[17] Stanley F Chen. Bayesian grammar induction for language modeling. *arXiv preprint cmp-lg/9504034*, 1995.

[18] Dan Klein and Christopher D Manning. A generative constituent-context model for improved grammar induction. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 128–135, 2002.

[19] Dan Klein and Christopher D Manning. Natural language grammar induction with a generative constituent-context model. *Pattern recognition*, 38(9):1407–1419, 2005.

[20] Dmitriy Genzel. Inducing a multilingual dictionary from a parallel multitext in related languages. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 875–882, 2005.

[21] Benjamin Snyder and Regina Barzilay. Unsupervised multilingual learning for morphological segmentation. In *Proceedings of acl-08: hlt*, pages 737–745, 2008.

[22] Benjamin Snyder, Tahira Naseem, Jacob Eisenstein, and Regina Barzilay. Unsupervised multilingual learning for pos tagging. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1041–1050, 2008.

[23] Tao Jiang, Lusheng Wang, and Kaizhong Zhang. Alignment of trees—an alternative to tree edit. *Theoretical Computer Science*, 143(1):137–148, 1995.

[24] Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394, 2010.

[25] Andriy Mnih and Geoffrey E Hinton. A scalable hierarchical distributed language model. *Advances in neural information processing systems*, 21:1081–1088, 2008.

[26] Jeremy Howard and Sebastian Ruder. Fine-tuned language models for text classification. *CoRR*, abs/1801.06146, 2018.

[27] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[28] Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. Learned in translation: Contextualized word vectors. In *Advances in neural information processing systems*, pages 6294–6305, 2017.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.