Intelligent Human Systems Integration (IHSI 2023), Vol. 69, 2023, 235–245
https://doi.org/10.54941/ahfe1002842

A Data Driven Approach for Automatic Language Acquisition

Ting Liu[1], Sharon Small[1], James Kubricht[2], Peter Tu[2], Harry Shen[1], Lydia Cartwright[1], and Samuil Orlioglu[1]

[1]Siena College, Loudonville, New York 12211, USA
[2]GE Global Research, Niskayuna, NY 12309, USA

ABSTRACT

Language acquisition has attracted more and more attention from researchers. Different from the traditional machine learning approaches, which require a lot of training data and are difficult to move to different domains, the current approaches in this area have started shifting gears to have machines to learn language like children, who create their own interpretations of the surrounding world through observation. Then, these interpretations will be mapped to common sense during interactions with their teachers/parents/friends. Furthermore, children are able to expand their knowledge dramatically through induction. In this paper, we present a novel data driven approach to simulate this process. The result shows that with the less effort of the human experts, the machine can learn knowledge faster and more productively.

Keywords: Language acquisition, Bootstrapping, Data driven, Clustering, Interactive learning

INTRODUCTION

Automatic Language Acquisition focuses on teaching an agent to acquire knowledge to understand the surrounding environment and be adaptive to a new environment. This is a process of passing human intelligence to agent, which is not a trivial task since people learn through generalization and induction. Research (Springer and Keil 1989, Keil 1992, Kelemen 1999 & 2003, Herrmann et al. 2010) in cognitive science has shown that children try to understand a new object from what they learned from other objects in the same category. The new findings will later be confirmed/fixed by teachers/parents/peers and then become the new knowledge. This self-motivated and interactive learning can significantly boost a child's knowledge with just a few examples.

However, traditional language understanding models: supervised, semi-supervised, and unsupervised, are very different from how children acquire knowledge. Supervised approaches (Emami and Jelinek 2005, Buys and Blunsom 2015, Dyer et al. 2016, Yin and Neubig 2017, Liu and Lapata 2018, Havrylov et al. 2019) are usually accurate but requires a large training dataset. It is expensive and time consuming, however, the trained model is difficult to shift to other domains. On the other hand, building unsupervised modals (Cai et al. 2018, Jin et al. 2018, Drozdov et al. 2019, Kim et al. 2019) is

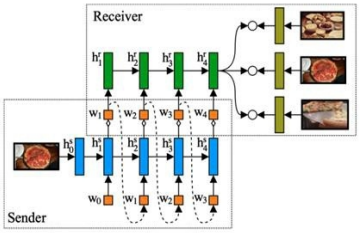

Figure 1: Architectures of sender and receiver LSTMs in Havrylov and Titov's (2017) EL implementation.

cheap and flexible, but its performance is usually significantly lower than those from the supervised approaches. With a relatively small set of guidance at the beginning, semi- supervised approach (Rei 2017, Rybak and Wróblewska 2018, Yin et al. 2018, Corro and Titov 2019, Zhu et al. 2020) can teach itself through the unlabelled dataset to achieve a comparable performance as a supervised modal. However, building the guidance is not a trivial task since the learning process won't be effective if the relationship between labelled data and unlabelled data is low.

Therefore, we proposed a data driven approach that puts an agent into the multimodal environment to learn knowledge through self-training and interaction with human expert. This has attracted researchers' attention in recent years (Matuszek et al. 2012, Antol et al. 2015, Chen et al. 2020). One branch aims to teach machine knowledge by simulating a children's learning process (Ross et al. 2018, Akimoto 2018). Even though the computing process is difficult, this is very attractive because it can help the agent accumulate knowledge from its experience and adapt itself into a new environment.

The following sections of this paper are: 1) The introduction of our system. 2) The data corpus we used for the learning process. 3) An unsupervised approach to extract an entity's attribute candidates from text. 4) Build knowledge through the interaction between the system and human experts on the candidates. 5) A bootstrapping process for finding more entities' attributes. 6) Evaluation. 7) Conclusion.

SYSTEM ARCHITECTURE

The ALAS (Automated Language Acquisition System) teaches the agent knowledge through both image/video and text. To describe the entities in an image/video, we utilized Havrylov and Titov's (2017 Figure 2) implementation as our working code base to generate 10 Emergent language (EL) codes (Kubricht et al. 2021), which are 3-digit numbers.

We then employed K-Means clustering to the EL codes to see how well they captured entities and their attributes. For example, clusters not only have a majority of members from one entity, e.g. a cluster entirely of horses, but also show a shared commonality, e.g., a cluster of only pear and apple. For each cluster, we built a Probabilistic Decision Tree (PDT) that represents which EL codes are best used to identify entities, their attributes (usually adjectives), and events (verbs or action nouns). These trees were saved as the knowledge

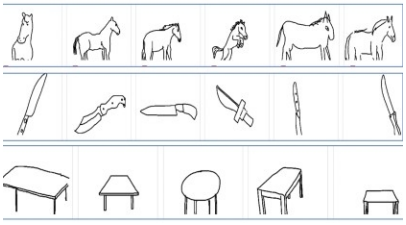

Figure 2: Images contained in the sketch dataset.

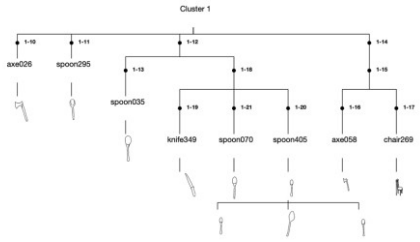

Figure 3: One cluster example generated from the images' EL codes.

of the agent in the Agent Knowledge Base (AKB). The details and evaluation of this process will be reported in a separate publication.

For the entities in each cluster, we employed our weighting method derived from tf-idf[1] to extract those important words from the text pieces (retrieved from a balance American English data corpus) containing the entities. Then the dialogue was triggered to ask human experts to assign the semantic relationships between the words and the entities if any exist. The input knowledge was then converted to patterns for the agent to find more words that have the similar relationships. The next sections will have a detailed explanation on how it works with the evaluation.

Data Set

We have collected 75,481 sketches of 125 different entities from Georgia Institute of Technology's Sketchy Database[2]. This collection has an average of 603.8 different sketches per entity, with a maximum of 746 and a minimum of 518 sketches per entity. We used this dataset to explore EL codes for the images. We argue that sketches capture many of the key attributes of an object class while simplifying the base representation from pixels to pen-strokes. Figure 2 provides samples from this dataset for three categories (horse, knife, and table). We taught the agent ten entities: apple, axe, chair, chicken, hammer, knife, owl, pear, spoon, and table, using the sketchy dataset. Figure 3 is an example PDT generated from a cluster using the images' EL codes. The majority portion of the cluster is spoon images. The other entities are knife and axe, which is reasonable since the image shape is quite similar to the spoons' and they all share the same attribute: handle.

[1] https://en.wikipedia.org/wiki/Tf%E2%80%93idf

[2] http://sketchy.eye.gatech.edu/

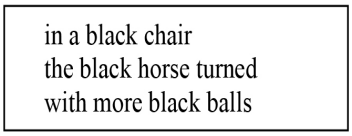

Figure 4: 4-gram examples.

Table 1. Distribution of three sample words in 4-grams containing apple.

Word	−3	−2	−1	1	2	3	Left	Right
Stem	0.1	0.05	0	0.38	0.33	0.14	0.15	0.85
Leaf	0.11	0.36	0	0.36	0.11	0.07	0.47	0.53
The	0.04	0.16	0.58	0.04	0.13	0.05	0.78	0.22

Another dataset we used is Corpus of Contemporary American English (COCA)[3]. As the "only large, genre-balanced corpus of American English" available online, COCA contains more than 560 million English words evenly divided among spoken, fiction, modern magazines, newspapers, and academic texts. This is a perfect dataset for us to get the common use of modern American English. Instead of using the whole corpus, we focused on 4-grams (Figure 4), which can either be a phrase or a short sentence, that contain the entities for the agent to learn.

Find Entities Correlated Words

For the 4-grams containing an entity, ALAS builds a context that contains 3 words before the entity and 3 words after, where we look for the words highly correlated to the entity. We employed the PDS approach (Liu et al, 2021) to the 4-grams containing the entity. For Example, this distribution of the words in context can be seen for the entity apple (Table 1). Then, we again utilized K-means clustering to build clusters that group the words that have the similar distribution. For example, the words, sauce and seed, are categorized together in one cluster where most words occur immediately after apple.

After clustering, we first used the medoids, the words closest to the center of a cluster, as the cluster's representation. The mid-column of Table 2 shows the percentage of the top 15 medoids that are semantically related to the entity apple. Only the result from cluster 4 is decent. The failure analysis showed that some of the top ranked words, like "until" and "whether", have no direct relationship with apple.[4] Therefore, we employed a method to calculate the word weight using the combination of word frequency[5] (wf)

[3]https://www.english-corpora.org/coca/

[4]We didn't add text preprocessing including POS tagging and syntactic parsing based on supervised machine learning technique since we plan to teach the machine to learning as a child learns, who doesn't need POS tagging to understand the meaning of words and the structures of sentences.

[5]How frequent a word occurs in the 4-grams of an entity. The higher the frequency, the more important the word to the entity.

Table 2. The accuracy of top 15 words in the clusters using two different ranking methods.

Cluster ID	Medoids	Weight
0	27%	33%
1	13%	40%
2	47%	100%
3	0	0
4	67%	73%

$$Weight(w,e) = wf(w,e) * ief(w,E)$$
$$wf(w,e) = frequency(w,e)$$
$$ief(w,E) = \log(\frac{N}{1+n(w)})$$

n(w) - number of entities have w in their 4-grams
N – total number of entities

Figure 5: The weight formula for the words in each cluster.

and inverted entity frequency[6] (ief) (Figure 5), which is a technique often used in information retrieval systems. The third column of Table 2 shows a significant improvement since a lot more top ranked words by weight are semantically related to apple. We then decided to use the weights to rank the words in a cluster.

In addition, we noticed that the number of relevant words from each cluster is very different from each other, which indicated that the words with certain distribution pattern(s) are highly semantically related with the entity. For example, the words occurring immediately before the entity are very likely the descriptive attributes, like color and size, of the entity. To verify the hypothesis, we first generated the clusters of all entities and assigned them into a category based on the distribution of the words in them. We defined six categories, i) Evenly distributed (E); ii) The second word on the left of the entity (1L); iii) The first word on the left (2L); iv) The first word on the right (1R); v) The second word on the right (2R); and vi) The third word on either left or right (3L/3R). Table 3 shows that the immediate left and right positions are the most frequent positions for the words that cooccurred with the entity and then followed by the second position on the right of the entity, which is explainable since most of the entity's attributes occurred in those positions.

The assessment (Table 4) of the clusters supports our hypothesis that the position of a word is a highly correlated with whether the word is semantically related to the entity when it occurs next to the entity. Interestingly, the clusters (from the 2nd column) with the words evenly distributed have

[6]How many entities in the dataset have a word in their 4-grams. The less entities that have the word, the more important the word is.

Table 3. The number of the clusters in each category.

	E	2L	1L	1R	2R	3L/3R
# of clusters	6	8	10	10	6	8

Table 4. The average of percentage of words in each category that are semantically related with the entity.

	E	2L	1L	1R	2R	3L/3R
Percentage	48.8%	20.9%	90%	66%	22.9%	26%

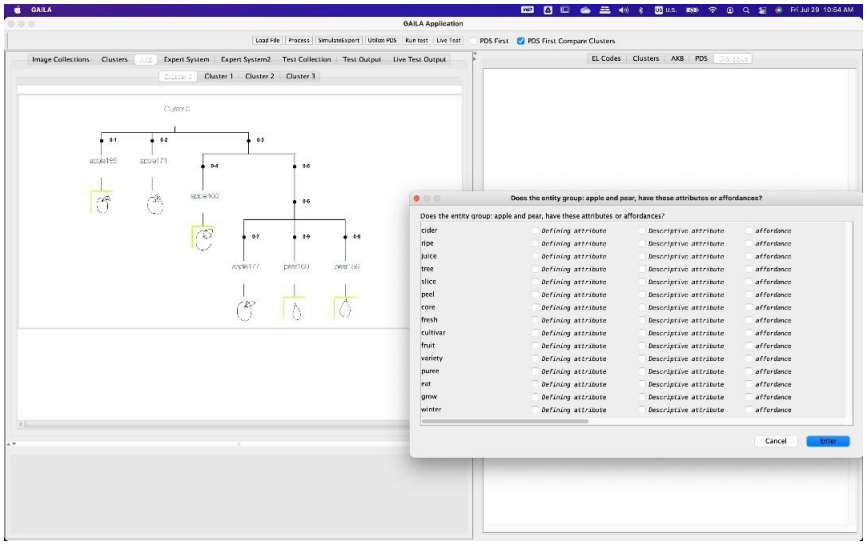

Figure 6: An example of the words given by ALAS for the entities in a cluster.

the third most semantic words because the clusters contain verbs that can be syntactically on both side of the entities.

From the above analysis, we found that there are three critical factors to pick proper candidates that are semantically correlated to an entity,

1. The tf-ief score of a word, the importance of the word to the entity.
2. The distribution category a word belongs to, the possibility a word has a semantic relationship with the entity.
3. The diversity of the candidates (from different distribution categories), the greater the diversity the candidates have, the more language patterns can be discovered.

Learn Entities' Attributes Given Human Experts Assessment

Currently, ALAS only uses tf-ief to select candidates and we are working on accommodating the other two features into the selection method. For the entities grouped by ALAS (Figure 3), the common top ranked candidates were selected and displayed to the expert for their input. Figure 8 shows an example with the top ranked common words, such as juice, cider, and eat, that occurred in the context of both apple and pear, which are the entities

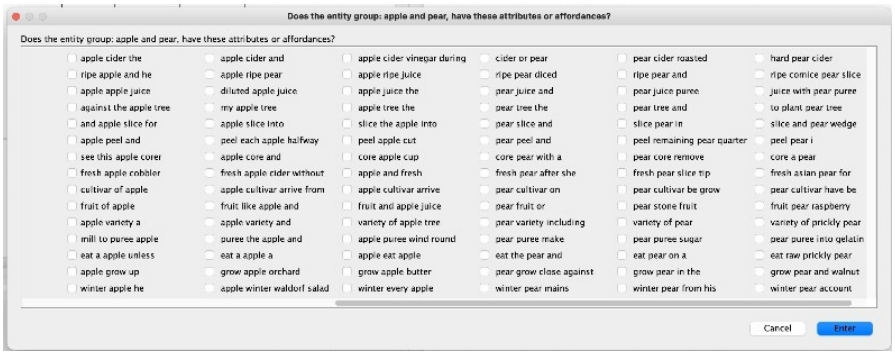

Figure 7: The most frequent 4-grams contains both candidates and the entity.

Table 5. Attributes found through the interaction with human experts.

	Defining	Descriptive	Affordance	Accepted by Expert	Total Suggestions
Number of attributes	22	32	57	111	240

in cluster one. The experts will mark whether a given word is an attribute of both apple and pear and if yes, what kind of attributes it is. We defined 3 attribute types for ALAS to learn: 1. Descriptive attributes: the attributes are used to describe the appearance of an entity, like size, color, and shape; 2. Defining attributes: the attributes are used to describe the components of an entity, like peel and core of apple and pear; 3. Affordance attributes: the attributes are used to describe how the entity can be used. For example, pear is a kind of food and apple can be used to make juice.

In addition to the word list given by ALAS, the system also displays the most frequent three 4-gram examples from each entity in the cluster. In this way, ALAS learns not only the entity's attributes but also how to use them. For example, in the 4-grams in Figure 7 containing pear and cider, "pear cider roasted" and "hard pear cider" are the positive examples that cider can be made from pear, but in the text piece, "cider or pear", there is no relationship between "cider" and "pear", therefore, this 4-gram is a negative example.

After experts go through all the suggestions provided by the PDS, ALAS will save the experts' inputs as new knowledge, which is applied for other entities to check whether it's in their top ranked candidate list. If yes, a dialogue will be triggered for confirmation. Figure 8 displays an example that "eat" is marked as an affordance attribute for both pear and apple and since it's in the top ranked word list of chicken and owl, ALAS then pops up a dialogue window and asks experts whether "eat" is also an attribute of chicken and owl.

Through the interaction, ALAS provided 240 words as the suggested attributes for the entities from eight clusters. Table 5 shows that 46.3% (111) of them were picked by the experts as entity attributes, which indicates that PDS suggestions are a good start. In addition to the learned knowledge, ALAS can also estimate the quality of the clusters based on the number of common

Figure 8: ALAS found that the affordance feature, eat, of apple and pear could also be the attribute of chicken and owl.

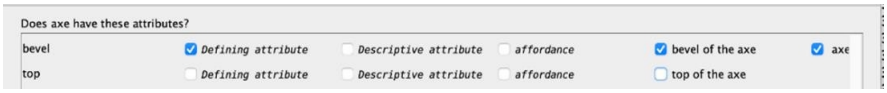

Figure 9: New feature candidates found for entity axe through the bootstrapping process.

attributes selected from experts. For example, experts picked 11 words as attributes from the cluster containing apple and pear but no attribute from the axe and chair cluster, which indicates that axe and chair don't share commonalities and shouldn't be clustered together.

Bootstrap More Attributes From the Learned Knowledge

Through the interaction with experts, ALAS learned not only the attributes but also how they are used in the text. Therefore, ALAS can bootstrap its knowledge by creating patterns from the 4-grams picked by the experts to find more attributes. To do so, we simply replaced the attribute words with star in the 4-grams. For example, the text, "pear skin and", that contains "skin" as the defining attribute of "pear" is transformed into "pear * and". Then for the similar patterns, e.g., "pear * and", "pear * on", and "pear * or", we did more generalization, "pear *". After all the patterns were generated, they were applied back to the 4-grams to extract more attribute candidates for experts to input. **Error! Reference source not found.**9 shows part of the pop-up window that contains new attribute's candidates for entity axe (instead of clusters) through the bootstrapping process. From this process, the system can learn not only more attributes but also the distinguished ones, such as the bevel of an axe, of one entity.

The bootstrapping process (Table 6) in total found 150 new candidates and 103 of them were picked by the experts. The accuracy is 68.7%, which is significantly better than the first suggestions by ALAS. It is reasonable since this learning step has human knowledge involved. Furthermore, the new patterns will be created from knowledge input by the experts in this step to find more features. With more and more knowledge collected, ALAS will have higher and higher confidence learning by itself and only need to interact with human experts occasionally to avoid introducing too much noise.

Table 6. New attributes found through PDS bootstrapping process.

	Defining	Descriptive	Affordance	Accepted by Expert	Total Suggestions
Number of attributes	19	66	15	103	150

In addition, we compared this data-driven approach with another ALAS model, which requires defining the entity's attributes ahead of time. In this model, we had four experts spend three hours to figure out what the defining attributes of the 10 entities were for ALAS to learn. After that, it took one expert two and half hours to teach ALAS 34 defining attributes of the ten entities. The data driven approach, on the other hand, is more efficient since it required only one hour for one expert to interact with ALAS on the initial suggestions and bootstrapped candidates. It is also much more productive since ALAS learned not only 41 defining attributes but also 98 descriptive and 72 affordance features through the two rounds learning process. Therefore, the data driven approach is a promising research direction and we are investing more time and efforts in pushing our system further in that direction.

CONCLUSION

In this paper, we proposed a novel data driven approach to learn the attributes of entities. It significantly reduced the human efforts in preparing the curriculum since the system gives the suggestions first. It also improved the learning quality significantly since it can bootstrap from the learned knowledge to discover more attributes of the entities. Here is the plan to improve the system,

1. We will find not only more entities' attributes but also new entities
2. We are optimizing the algorithm in selecting the best candidates to improve the learning efficiency
3. We are looking for the frequency to involve experts in the learning process after the first few iteration to guard the learning quality.

ACKNOWLEDGMENT

This research was funded by the DARPA GAILA program. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

REFERENCES

Akimoto, T. (2018). Stories as mental representations of an agent's subjective world: A structural overview. Biologically Inspired Cognitive Architectures, 25, 107–112.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C L., and Parikh, D. (2015). VQA: Visual question answering. *In Proceedings of the IEEE International Conference on Computer Vision*. pages 2425–2433.

Buys, J., and Blunsom, P. (2015). Generative Incremental Dependency Parsing with Neural Networks. *In Proceedings of ACL*.

Cai, J.; Jiang, Y.; and Tu, K. (2017). CRF Autoencoder for Unsupervised Dependency Parsing. *In Proceedings of EMNLP*.

Chen, Z.; Wang, P.; Ma, L.; Wong, K.-Y. K.; and Wu, Q. (2020). Cops-Ref: A new Dataset and Task on Compositional Referring Expression Comprehension. *CoRR abs/2001.09308*.

Corro, C., and Titov, I. (2019). Differentiable Perturb and Parse: Semi-Supervised Parsing with a Structured Variational Autoencoder. *In Proceedings of ICLR*.

Drozdov, A.; Verga, P.; Yadev, M.; Iyyer, M.; and McCallum, A. (2019). Unsupervised Latent Tree Induction with Deep Inside-Outside Recursive AutoEncoders. *In Proceedings of NAACL*.

Dyer, C., Kuncoro, A., Ballesteros, M., and Smith, N. A. (2016). Recurrent Neural Network Grammars. *In Proceedings of NAACL*.

Emami, A., Jelinek, F. (2005). A Neural Syntactic Language Model. *Mach Learn* 60, 195–227.

Havrylov, S.; Kruszewski, G.; and Joulin, A. (2019). Cooperative Learning of Disjoint Syntax and Semantics. *In Proceedings of NAACL*.

Herrmann, P, Waxman, SR, Medin, DL. (2010). Anthropocentrism is not the first step in children's reasoning about the natural world. Proc Natl Acad Sci USA. 2010 Jun 1;107(22):9979–84. doi: 10.1073/pnas.1004440107.

Jin, L.; Doshi-Velez, F.; Miller, T.; Schuler, W.; and Schwartz, L. (2018). Unsupervised Grammar Induction with Depth-bounded PCFG. *In Proceedings of TACL*.

Keil, F. C. (1992). The origins of autonomous biology. In M. R. Gunnan & M. Maratsos (Eds.), Minnesota symposium on child psychology: Modularity and constraints on language and cognition (pp. 103–137). Hillsdale, NJ: Erl- baum.

Kelemen, D. (1999). Function, goals and intention: Children's teleological reasoning about objects. Trends Cogn. Sci. 1999;3:461–468. doi: 10.1016/S1364-6613(99)01402-3.

Kelemen, D. (2003). British and American children's preferences for teleo-functional explanations of the natural world. Cognition, 88, 201–221.

Kim, Y.; Rush, A. M.; Yu, L.; Kuncoro, A.; Dyer, C.; and Melis, G. (2019). Unsupervised Recurrent Neural Network Grammars. *In Proceedings of NAACL*.

Kubricht, J., Small, S., Liu, T., and Tu, P. (2021). Towards an automated language acquisition system for grounded agency. In Proceedings of SAI Intelligent Systems Conference, pages 23–43. Springer, 2021.

Liu, T., Small, S, Kubricht, J., and Tu, P. (2021) A Semi-Supervised Machine Learning Method for Language Acquisition. In proceedings of AAAI Hybrid Artificial Intelligence Workshop.

Liu, Y., and Lapata, M. (2018). Learning Structured Text Representations. *In Proceedings of TACL*.

Matuszek, C.; FitzGerald, N.; Zettlemoyer, L.; Bo, L.; and Fox, D. (2012). A joint model of language and perception for grounded attribute learning. *ICML*.

Rei, M. (2017). Semi-supervised Multitask Learning for Sequence Labeling. *In Proceedings of ACL*.

Ross, C.; Barbu, A.; Berzak, Y.; Myanganbayar, B.; and Katz, B. (2018). Grounding language acquisition by training semantic parsers using captioned videos. *In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2647–2656.

Rybak, P., and Wróblewska, A. (2018). Semi-Supervised Neural System for Tagging, Parsing and Lematization. *In Proceedings of CoNLL*.

Springer, K., & Keil, F. C. (1989). On the development of biologically specific beliefs: The case of inheritance. Child Development, 60(3), 637–648.

Yin, P. and Neubig, G. (2017). A Syntactic Neural Model for General-Purpose Code Generation. *In Proceedings of ACL*.

Yin, P.; Zhou, C.; He, J.; and Neubig, G. (2018). StructVAE: Tree-structured Latent Variable Models for Semi-supervised Semantic Parsing. *In Proceedings of ACL*.

Zhu, S.; Cao, R.; and Yu, K. (2020). Dual Learning for Semi-Supervised Natural Language Understanding. *IEEE/ACM Transactions on Audio Speech and Language Processing*.