Anaphoric Zero Pronoun Identification: A Multilingual Approach

Abdulrahman Aloraini[1,2], **Massimo Poesio**[1]
[1]School of Electronic Engineering and Computer Science, Queen Mary University of London
[2]Department of Information Technology, Qassim University
{a.aloraini, m.poesio}@qmul.ac.uk

Abstract

Pro-drop languages such as Arabic, Chinese, Italian or Japanese allow morphologically null but referential arguments in certain syntactic positions, called anaphoric zero-pronouns. Much NLP work on anaphoric zero-pronouns (AZP) is based on gold mentions, but models for their identification are a fundamental prerequisite for their resolution in real-life applications. Such identification requires complex language understanding and knowledge of real-world entities. Transfer learning models, such as BERT, have recently shown to learn surface, syntactic, and semantic information,which can be very useful in recognizing AZPs. We propose a BERT-based multilingual model for AZP identification from predicted zero pronoun positions, and evaluate it on the Arabic and Chinese portions of OntoNotes 5.0. As far as we know, this is the first neural network model of AZP identification for Arabic; and our approach outperforms the state-of-the-art for Chinese. Experiment results suggest that BERT implicitly encode information about AZPs through their surrounding context.

1 Introduction

Empty categories provide an important source of syntactic information about the phonetically null arguments in pro-drop languages such as Arabic (Eid, 1983), Chinese (Li and Thompson, 1979), Italian (Di Eugenio, 1990), Japanese (Kameyama, 1985), and others (Bever and Sanz, 1997; Kim, 2000). The use of empty categories started with Penn Treebanks (Marcus et al., 1993), followed by Arabic Treebank (Maamouri et al., 2004), Chinese Treebank (Xue et al., 2005) and other Penn-style series. Empty categories are used to represent traces, such as, movement operations in interrogative sentence, also to represent right node raising which is a shared argument in the rightmost constituent of a coordinate structure. Another usage of empty categories is zero-pronouns (ZP) which are omitted pronouns in places where they are expected to be, and function as overt pronouns. Anaphoric zero pronouns (AZP) are ZPs that corefer to one or more noun phrases in a preceding text. The following example of an AZP comes from the Arabic section of OntoNotes:

.. المفارقة الأخرى عن بوش هي عدم حماسته للمؤتمر الدولي، اذ انه من البداية، يريد * اجتماعا مختلفا ....

Ironically, Bush did not show any enthusiasm for the international conference, because since the beginning, (he) wanted to attend another conference ...

In the example, the ZP indicated with '*' refers to the gap position of an omitted pronoun (In OntoNotes 5.0, ZPs are denoted as ***** in Arabic text, and ***pro*** in Chinese). The omitted pronoun refers to a singular masculine person that has been mentioned previously, in the example "Bush/بوش". In Arabic, we deduce the reference information from the context, especially the verb that precedes the AZP, in the example the verb is "wanted/يريد". Since English is not a pro-drop language (White, 1985), the AZP gap position is translated into an overt pronoun (he). The AZP problem has inspired much research because it benefits many natural language processing tasks such as machine translation (Mitkov and Schmidt, 1998), and coreference resolution (Mitkov et al., 2000). Recently, there has been a great deal of research on AZPs for Chinese (Kong et al., 2019; Yin et al., 2018; Chang et al., 2017; Liu et al., 2017; Yin et al., 2017), Arabic (Aloraini and Poesio, 2020), Japanese (Shimazu et al., 2020), Korean (Jung and Lee, 2018), and other languages (Grigorova, 2016; Gopal and Jha, 2017). A major drawback of many existing studies is the assumption that AZP locations are given; hence, they focus primarily on resolving AZPs to their correct antecedent. However, such assumption does not reflect real-life applications. Another drawback is that current AZP identification systems rely on language-dependent features and fail to detect many AZP locations. In addition, some languages do not have an AZP identification system, one of which is Arabic.

Proceedings of the 3rd Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2020), pages 22–32, Barcelona, Spain (online), December 12, 2020.

To alleviate the above-mentioned limitations, we investigate the AZP identification task and study if the recently achieved state-of-the-art transfer learning methods, such as BERT (Devlin et al., 2018), can work well on identifying AZPs. Typically, AZP identification task consists of two steps. The first is the *extraction* step where potential ZP locations are extracted. The extraction procedure is based on heuristics and depend on the target language structure. The second step is *classification* step which determines which of the extracted candidate are AZP. The classification step is more challenging because of the varieties and size of the extracted candidates. In this paper, we propose a multilingual approach to AZP identification based on BERT. We make three main contributions:

- We propose a BERT-based multilingual model and evaluate on languages that differ completely in their morphological structure: Arabic and Chinese. (Arabic is morphologically rich, whereas Chinese's morphology is relatively simple (Pradhan et al., 2012))
- Ours is the first neural network-based AZP identification model for Arabic, and it substantially surpasses the current state-of-the-art on Chinese.
- Our experimental results suggest that BERT representations encode information about AZPs through their context.

The rest of the paper is organized as follows. We review Arabic and Chinese ZP-related literature, and other languages as well in Section 2. We explain our proposed model in Section 3. We discuss the evaluation settings in Section 4. We show the results and discuss them in Section 5. We conclude in Section 6.

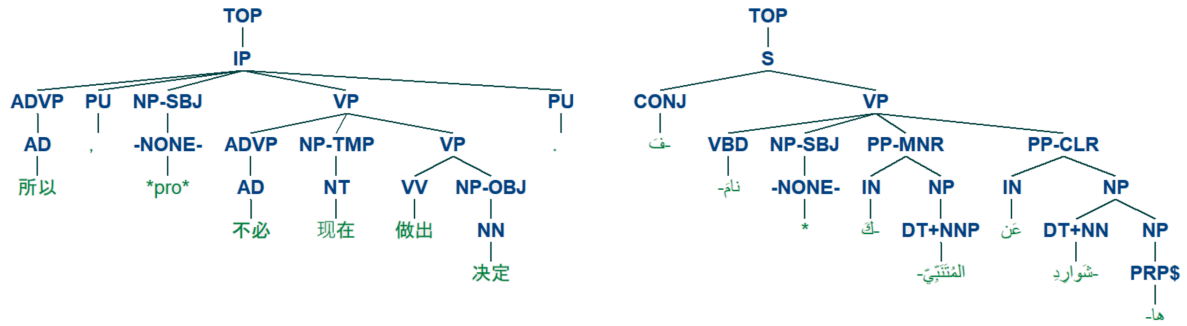

Figure 1: Chinese ZPs appear before a VP node (left), and Arabic ZPs appear after the verb of a VP head (right).In OntoNotes 5.0, Chinese AZPs are annotated as ***pro*** and Arabic AZPs as *.

2 Related Work

AZP identification task has been considered independently, but also as a prerequisite step before AZP resolution task because the detection has a heavy impact on the resolution (Kong et al., 2019).

Arabic: There have been a few studies devoted to AZPs and empty categories in general. Green et al. (2009) proposed a conditional-random-field (CRF) sequence classifier to detect Arabic noun phrases, and captured ZPs implicitly. Bakr et al. (2009) applied a statistical approach to detect empty categories. Gabbard (2010) proposed a pipeline made of maximum entropy classifiers which jointly make a CRF to retrieve Arabic empty categories. Aloraini and Poesio (2020) proposed the first neural model for resolving Arabic AZP, but they did not consider the AZP identification step. As far as we know, no previous work has considered Arabic AZP identification.

Chinese: Converse (2006) studied AZP resolution and applied a rule-based approach that employed Hobbs algorithm (Hobbs, 1978) to resolve ZPs in the Chinese Treebank; however, did not attempt to automatically identify AZP. Yeh and Chen (2006) is another rule-based approach, for AZP resolution and also used a set of hand-engineered rules to identify AZPs. Zhao and Ng (2007), the first machine learning approach to Chinese AZPs identification and resolution, by applying decision trees incorporated with a set of syntactic and positional features. (Kong and Zhou, 2010) employed a tree kernel-based approach to AZP identification and resolution. Chen and Ng (2013) is an extension of (Zhao and Ng, 2007), they incorporated contextual features for AZP resolution and applied a combination of syntactic, lexical and other features for the identification. Chen and Ng (2014) proposed unsupervised techniques to resolve AZPs and applied a set of rules to identify AZP. Chen and Ng (2015) is another unsupervised approach on the AZP resolution. Recent approaches applying deep-learning neural networks include Chen and Ng (2016) trained a binary classifier to identify AZP and applied a feed-forward neural network to the AZP resolution; Yin et al. (2016) used (Chen and Ng, 2016)'s classifier to identify AZPs. For AZP resolution, they employed an LSTM to represent AZP and two subnetworks (general encoder and local encoder) to capture context-level and word-level information of the candidates; Yin et al. (2017) also applied (Chen and Ng, 2016)'s classifier

to detect AZPs and proposed an improved deep memory network to resolve AZPs; and Liu et al. (2017), applied an attention-based neural network to resolve AZPs and enhanced the performance by training on automatically generated large-scale training data. Chang et al. (2017) focused primarily on AZP identification and applied an LSTM neural-network with text and part-of-speech information. Yin et al. (2018), also used an attention-based model, but combined their network with (Chen and Ng, 2016) features to resolve AZPs. Yin et al. (2019) applied the same heuristics in (Chen and Ng, 2015) to identify AZPs and applied a collaborative-filtering approach to resolve AZPs. Kong et al. (2019) identified AZPs using a learning-based classifier with semantic, lexical and syntactic features, and used coreferential chain information to improve AZP resolution.

Other languages: There has been also a great deal of research on identification and resolution of AZPs, particularly in Japanese (Yoshimoto, 1988; Kim and Ehara., 1995; Aone and Bennett, 1995; Seki et al., 2002; Isozaki and Hirao, 2003; Iida et al., 2006; Iida et al., 2007; Sasano et al., 2008; Sasano et al., 2009; Sasano and Kurohashi, 2011; Yoshikawa et al., 2011; Hangyo et al., 2013; Iida et al., 2015; Yoshino et al., 2013; Yamashiro et al., 2018), but also in other languages, including Korean (Han, 2004; Byron et al., 2006), Spanish (Ferrández and Peral, 2000; Rello and Ilisei, 2009), Portuguese (Rello et al., 2012), Romanian (Mihăilă et al., 2011), Bulgarian (Grigorova, 2013), and Sanskrit (Gopal and Jha, 2017). Iida and Poesio (2011) proposed the first multilingual approach for AZP resolution.

Current approaches suffer from one (or more) of the following. First, they assume AZPs are available; so they focus mainly on the resolution part. Second, they apply on a private or very small size corpus. Third, they rely on an extensive set of features or language-dependent rules to identify AZP.

3 Model

To identify AZPs, context understanding and semantic knowledge of entities are essential in Chinese (Huang, 1984) as well as in Arabic which requires, in addition, deep understanding of its complex morphology (Alnajadat, 2017). Recently, it has been shown that BERT (Devlin et al., 2018) can capture structural properties of a language, such as its surface, semantic, and syntactic aspects (Jawahar et al., 2019) which seems suitable for the AZP identification task. Therefore, we use BERT to produce representations for ZP candidates. Our model is a binary classifier that takes an automatically predicted ZP candidate as input, and classifies it as an AZP or not. In this section, we first give an overview of BERT and its adaptation modes. We then describe how we generate AZP candidates, and how we represent them. Finally, we present the training objective and hyperparameter tuning settings.

3.1 BERT

BERT is a language representation model consisting of multiple stacked Transformers (Vaswani et al., 2017). BERT was pretrained on a large amount of unlabeled text, and produces distributional vectors for words and contexts. BERT was pretrained on different settings, we use BERT-base Multilingual which was pretrained on many languages, including Chinese and Arabic, and is publicly available[1]. BERT has two modes of adaptation: feature extraction and fine-tuning. Feature extraction (also called feature-based) is when BERT representations are used as they were originally pretrained, without any further training. Fine-tuning is the process of slightly adjusting BERT's parameters for a target task. Feature extraction is computationally cheaper and might be more suitable for a specific task (Peters et al., 2019). Fine-tuning is more convenient to utilize, but restricted to several general-purpose tasks. AZP identification task was not pretrained as part of BERT tasks and not directly applicable to fine tuning mode without any modifications to BERT's architecture. We employ feature extraction mode to represent AZP candidate in our classifier.

3.2 Candidate Generation

Although ZPs are annotated in OntoNotes, our model works off automatically predicted candidates. ZP locations differ in Chinese and Arabic. In Chinese, ZPs appear before a VP node while in Arabic they appear after the head of a VP node [2]. An example of Chinese and Arabic ZP locations in Figure 1. We extract Chinese ZP locations as in (Zhao and Ng, 2007)'s work. They consider every gap before a VP node as a candidate. The number of candidates can be large. (Kong and Zhou, 2010) showed that if a VP node is in a coordinate structure or modified by an adverbial node, only its parent VP node is considered, thus decreasing the number of necessary candidates. For Arabic, we consider every gap after every head of a VP node as a candidate. A candidate is positive if it is an AZP, negative otherwise. Both approaches result in extracting many negatives examples and a small number of positive examples. The high imbalance between the two classes can make a model biased; we address the problem in Section 5.

[1] https://storage.googleapis.com/bert_models/2018_11_23/multi_cased_L-12_H-768_A-12.zip

[2] There are two types of word order for Arabic: Subject-Verb-Object and Verb-Subject-Object. Both are used and acceptable. In the annotation process, Arabic Treebank sets the Verb-Subject-Object as the official order.

3.3 Input Representation

We represent AZPs by their surrounding context, specifically, we represent each candidate by its VP headword and its context window of two words (left and right). Consider a sentence with a gap candidate *C* at position *i*, so its surrounding context at positions *i-2*, *i-1*, *i+1*, *i+2*.

$$sentence = (w_1, w_2, ...w_{i-2}, w_{i-1}, C_i, w_{i+1}, w_{i+2}, .., w_n) \tag{1}$$

We feed *sentence* into BERT feature extraction mode as input and it outputs *embeddings* of every word of *sentence*.

$$embeddings = BERT(sentence) \tag{2}$$

We extract the embeddings of the candidate position and its surrounding context. In our experiments, BERT Tokenizer, Wordpiece (Wu et al., 2016), segmented many Arabic words into multiple sub-tokens, each with its own embeddings. For example, the word *sleeping* might be segmented into two sub-tokens *sleep* and *##ing*. One way to represent word sub-tokens is to compute their mean; therefore, we create the function μ which computes the mean of sub-token embeddings. We join the AZP context representations together into a value called *azp*.

$$a_1 = \mu(embeddings_{(i-2)}) \tag{3}$$
$$a_2 = \mu(embeddings_{(i-1)}) \tag{4}$$
$$a_3 = \mu(embeddings_{(i)}) \tag{5}$$
$$a_4 = \mu(embeddings_{(i+1)}) \tag{6}$$
$$a_5 = \mu(embeddings_{(i+2)}) \tag{7}$$
$$azp = [a_1, a_2, a_3, a_4, a_5] \tag{8}$$

azp encodes information about the candidate context and serves as input to our classifier. It is possible to extend the AZP window to more context but we empirically find context window of size 2 to be sufficiently effective.

$$layer_1 = f(W_1 azp + b_1) \tag{9}$$
$$layer_2 = f(W_2\, layer_1 + b_2) \tag{10}$$
$$output = f(W_3\, layer_2 + b_3) \tag{11}$$

The binary classifier is a multi-layered perceptrons consisting of two layers and one output layer. f is the ReLU activation function (Nair and Hinton, 2010). Each layer in the classifier has learning parameters W and b. The input is then classified to be either an AZP or not.

3.4 Training Objective

The training objective of our classifier is binary cross-entropy loss:

$$J(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i)\right] \tag{12}$$

θ represents the set of learning parameters in the model. N is the number of training data. y_i is the true label of training *i* and $\hat{y}_i$ its predicted label.

3.5 Hyperparameter Tuning

In the classifier, we employ two layers and initialize each one's weights using Glorot and Bengio (2010)'s method. We also add a dropout regularization between the two layers and the output layer. We tune the hyperparameters based on the development sets. Table 1 shows the hyperparameter settings.

4 Evaluation

4.1 Datasets

We evaluate our model on the Arabic and Chinese subsets of OntoNotes 5.0, which were used in the the official CoNLL-2012 shared task (Pradhan et al., 2012).

Chinese training and development sets contain AZPs, but the test set does not. Therefore, we train the model using the training set and we use the development set as the test set, a practice followed in prior research (Chen and Ng, 2013; Chen and Ng, 2014; Chen and Ng, 2016; Kong et al., 2019). We reserve 20% of the training data as a

Number of units in the first layer	800
Number of units in the second layer	600
Number of training epochs	10
Learning rate	1e-5
Dropout rate	0.5
Optimizer	Adam

Table 1: Hyperparameter settings.

Language	Category	Training	Dev	Test
Chinese	Documents	1,391	172	N/A
	Sentences	36,487	6,083	
	Words	756,063	100,034	
	AZPs	12,111	1,713	
Arabic	Documents	359	44	44
	Sentences	7,422	950	1,003
	Words	264,589	30,942	30,935
	AZPs	3,495	474	412

Table 2: Statistics on Chinese and Arabic datasets. Chinese test portion does not contain AZPs; therefore, the development portion is used for evaluation.

development set.

Arabic training, development, and test sets all have AZPs, and we use each set for its purpose. We preprocessed Arabic text by normalizing all variants of the letter "alif" and also removing all diacritics.

Detailed statistics about Chinese and Arabic dataset can be found in Table 2.

4.2 Metrics

We evaluate the results in terms of recall, precision, and F-score, as defined in (Zhao and Ng, 2007):

$$Recall = \frac{AZP\,hits}{Number\ of\ AZPs\ in\ Key}$$

$$Precision = \frac{AZP\,hits}{Number\ of\ AZPs\ in\ Response}$$

Key represents the true set of AZP entities in the dataset, and *Response* represents the system output of the identified AZPs in the model. *AZP hits* are the reported AZP positions in *Response* which occur in the same position as in *Key*.

5 Results

AZP identification results for Arabic are in Table 3, and Chinese in Table 4. The training data is highly imbalanced because of the ratio of negatives examples to the positive examples. In Arabic there are 5.6 times of negative examples compared to the positive examples, and in Chinese the negative examples are 16.2 times compared to the positive ones. To address this problem, we follow (Zhao and Ng, 2007)'s approach by changing the ratio weight *r* of sampling positive examples with respect to negative examples. The value *r* affects precision and recall scores. If *r* is high, precision increases but recall decreases. The effect of tuning *r* on precision, recall and F1 scores on Arabic and Chinese are in Figures 2 and 3 respectively. F1 scores with different variations of *r* are not very significant; however, we choose *r* that balances between the precision and recall scores.

Prior works (Chen and Ng, 2013; Chen and Ng, 2014; Chen and Ng, 2016; Chang et al., 2017; Kong et al., 2019) evaluate AZP identification under two settings: gold and system parse because annotation quality can impact the number of recovering candidates in the *extraction* step. Gold annotations are available for both languages and we also automatically parse the data with syntactic trees using the Berkeley Parser (Kitaev et al., 2018) which is a pre-trained parser using neural networks and self-attention.

5.1 Arabic

As far as we know, there has been no published proposal on Arabic AZP identification. Therefore, we implemented as a baseline (Chang et al., 2017)'s model, which employs sentence and Part-of-Speech information into a Bi-LSTM neural network to identify ZPs. We set its embedding layer to the Arabic version of Fasttext (Bojanowski et al.,

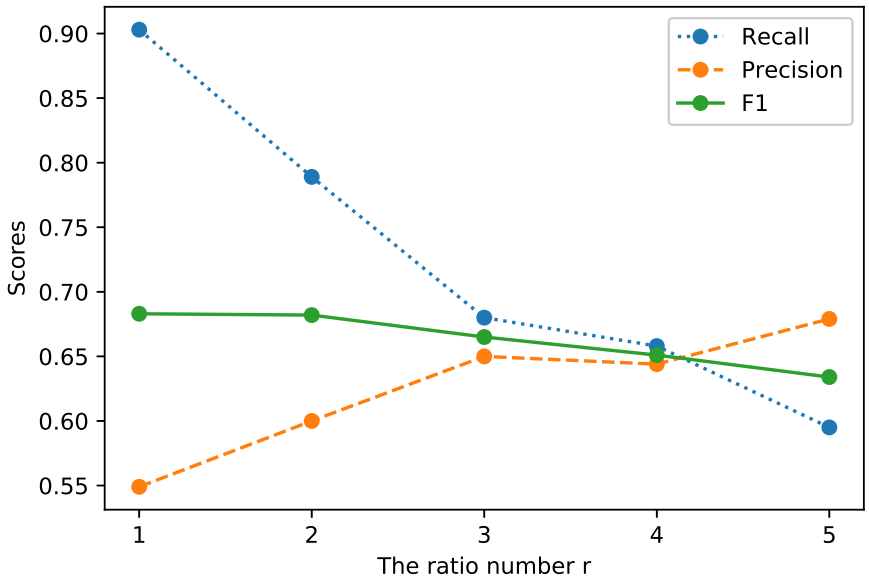

Figure 2: The effect of tuning the ratio *r* on recall, precision and F1 scores on the Arabic test set.

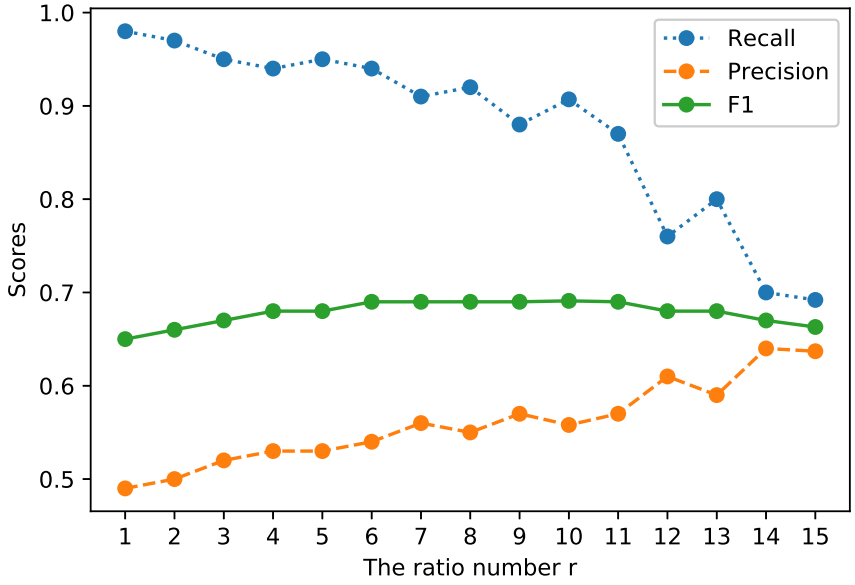

Figure 3: The effect of tuning the ratio *r* on recall, precision and F1 scores on the Chinese test set.

2017). We can see in Table 3 that our approach outperforms the baseline in both gold and system settings with F1 scores of 68.2% and 47.0%. There is a a big gap between gold and system parse because the automatic parser failed to recognize many VP nodes in the *extraction* step. Thus, many AZP samples were not recognized for training and evaluation which lead to a great decrease in performance. To gain additional insights into our model, we analyzed its outputs. The model correctly identifies many AZP cases, however, it struggles to recognize some patterns especially AZPs that are preceded by a verb in the first person. The errors can be attributed to the distribution of the training data. Most training AZP data are headed by verbs in the third person, and the number of verbs in the first and second persons is very small; thus, the model did not learn to classify many of these cases. A corpus that include a larger distribution of such cases can help a model to learn them.

	Settings 1: Gold Parse			Settings 2: System Parse		
	R	P	F1	R	P	F1
Baseline	**67.7**	45.2	54.2	31.7	30.6	31.1
Our model (r=2)	60.0	**78.9**	**68.2**	**38.6**	**60.1**	**47.0**

Table 3: AZP identification results for Arabic. The highest score is in **bold**.

5.2 Chinese

We compare our approach with other proposals in Table 4. As we can see, our approach achieves the highest F1 scores of 69.1% and 68.7% with gold and system parse settings, outperforming all prior proposals. The F1-score difference between our approach and the state-of-the-art approach is 4.7% with gold parse settings and 11.3% with system parse. The F1-score difference of gold and system settings of our approach is relatively small (0.4%) because the Berkeley parser annotated many VP nodes correctly. We analyzed the errors, and noticed many unidentified AZPs are located at the beginning of their samples. These cases depend on previous sentences, and their information might have not been encoded in the AZP input; thus, our model failed to identify them.

	Settings 1: Gold Parse			Settings 2: System Parse		
	R	P	F1	R	P	F1
(Chen and Ng, 2013)	50.6	55.1	52.8	30.8	34.4	32.5
(Chen and Ng, 2014)	72.4	42.3	53.4	42.3	26.8	32.8
(Chen and Ng, 2016)	75.1	50.1	60.1	43.7	30.7	36.1
(Chang et al., 2017)	63.5	**65.3**	64.4	57.2	55.7	56.4
(Kong et al., 2019)	70.1	59.4	64.3	60.2	40.2	48.2
Our model (r=10)	**90.7**	55.8	**69.1**	**81.9**	**59.2**	**68.7**

Table 4: AZP identification results for Chinese. The highest score is in **bold**.

5.3 Discussion

BERT representations work interestingly well on AZPs even though empty categories have not been considered during the BERT's pretraining. Recent works (Jawahar et al., 2019; Kovaleva et al., 2019; Goldberg, 2019; Clark et al., 2019) have shown that BERT learns various linguistic information such as, syntactic roles, coreference resolution, semantic relations and others. Our experimental results suggest that these information might be encoded in AZP contexts which make them distinctive.

Current approaches for AZP identification evaluate under two settings: gold and system annotations because the task depend highly on the annotation quality of parse trees. In our experiments, gold settings for both Arabic and Chinese achieve outstanding results. In system parse, Chinese achieves results similar to its gold setting; however, Arabic does not. The reason is that Berkeley Parser (Kitaev et al., 2018) fails to parse correctly Arabic sentences which means many correct AZP locations are not detected in the extraction step. A sophisticated Arabic parser can improve the overall performance for system-parse settings.

6 Conclusion

We proposed a BERT-based model for AZP identification. Our approach is multilingual, and we evaluate on Arabic and Chinese portions of OntoNotes. The model is the first to deal with Arabic AZP identification and the experiments demonstrated that our method surpasses the state-of-the-art on Chinese AZPs. In addition, our experimental results show that BERT learn about anaphoric zero-pronouns through their surrounding context.

References

Bashir M. Alnajadat. 2017. Pro-drop in standard arabic. In *International Journal of English Linguistics 7.1*.

Abdulrahman Aloraini and Massimo Poesio. 2020. Cross-lingual zero pronoun resolution. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 90–98.

Chinatsu Aone and Scott William Bennett. 1995. Evaluating automated and manual acquisition of anaphora resolution strategies. In *ACL '95 Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 122–129.

Hitham M Abo Bakr, Khaled Shaalan, and Ibrahim Ziedan. 2009. A statistical method for detecting the arabic empty category. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt*.

Thomas G Bever and Montserrat Sanz. 1997. Empty categories access their antecedents during comprehension: Unaccusatives in spanish. *Linguistic Inquiry*, pages 69–91.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Donna K. Byron, Whitney Gegg-Harrison, and Sun-Hee Lee. 2006. Resolving zero anaphors and pronouns in korean. In *Traitement Automatique des Langues 46.1*, pages 91–114.

Tao Chang, Shaohe Lv, Xiaodong Wang, and Dong Wang. 2017. Zero pronoun identification in chinese language with deep neural networks. In *2017 2nd International Conference on Control, Automation and Artificial Intelligence (CAAI 2017)*. Atlantis Press.

Chen Chen and Vincent Ng. 2013. Chinese zero pronoun resolution: Some recent advances. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1360–1365.

Chen Chen and Vincent Ng. 2014. Chinese zero pronoun resolution: An unsupervised probabilistic model rivaling supervised resolvers. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.

Chen Chen and Vincent Ng. 2015. Chinese zero pronoun resolution: A joint unsupervised discourse-aware model rivaling state-of-the-art resolvers. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 320–326.

Chen Chen and Vincent Ng. 2016. Chinese zero pronoun resolution with deep neural network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

Susan Converse. 2006. Pronominal anaphora resolution in chinese. In *PhD Thesis, University of Pennsylvania*.

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

B. Di Eugenio. 1990. Centering theory and the italian pronominal system. In *Proc. of the 13th COLING*, Helsinki, Finland.

Mushira Eid. 1983. On the communicative function of subject pronouns in arabic. In *Journal of Linguistics 19.2*, pages 287–303.

Antonio Ferrández and Jesús Peral. 2000. A computational approach to zero-pronouns in spanish. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 166–172.

Ryan Gabbard. 2010. Null element restoration. In *Ph.D Thesis, University of Pennsylvania*.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. In *arXiv preprint arXiv:1901.05287*.

Madhav Gopal and Girish Nath Jha. 2017. Zero pronouns and their resolution in sanskrit texts. In *The International Symposium on Intelligent Systems Technologies and Application*, pages 255–267.

Spence Green, Conal Sathi, and Christopher Manning. 2009. Np subject detection in verb-initial arabic clauses. In *Proceedings of the Third Workshop on Computational Approaches to Arabic Script-based Languages (CAASL3). Vol. 112.*

Diana Grigorova. 2013. An algorithm for zero pronoun resolution in bulgarian. In *Proceedings of the 14th International Conference on Computer Systems and Technologies*.

Diana Grigorova. 2016. Hybrid approach to zero pronoun resolution in bulgarian. In *Proceedings of the 17th International Conference on Computer Systems and Technologies 2016*, pages 331–338.

Na-Rae Han. 2004. A korean null pronouns: Classification and annotation. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation. Association for Computational Linguistics, 2004.*, pages 33–40.

Masatsugu Hangyo, Daisuke Kawahara, and Sadao Kurohashi. 2013. Japanese zero reference resolution considering exophora and author/reader mentions. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 924–934.

Jerry Hobbs. 1978. Resolving pronoun references. In *Lingua*, pages 311–338.

C.-T. James Huang. 1984. On the distribution and reference of empty pronouns. In *Linguistic Inquiry, Vol. 15, No. 4*, pages 531–574.

Ryu Iida and Massimo Poesio. 2011. A cross-lingual ilp solution to zero anaphora resolution. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 804–813.

Ryu Iida, Kentaro Inui, and Yuji Matsumoto. 2006. Exploiting syntactic patterns as clues in zero-anaphora resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistic*, pages 625–632.

Ryu Iida, Kentaro Inui, and Yuji Matsumoto. 2007. Zero-anaphora resolution by learning rich syntactic pattern features. In *ACM Transactions on Asian Language Information Processing, 6(4)*.

Ryu Iida, Kentaro Torisawa, Chikara Hashimoto, Jong-Hoon Oh, and Julien Kloetzer. 2015. Intra-sentential zero anaphora resolution using subject sharing recognition. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2179–2189.

Hideki Isozaki and Tsutomu Hirao. 2003. Japanese zero pronoun resolution based on ranking rules and machine learning. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing,*, pages 184–191.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy*.

Sangkeun Jung and Changki Lee. 2018. Deep neural architecture for recovering dropped pronouns in korean. *ETRI Journal*, 40(2):257–265.

Megumi Kameyama. 1985. *Zero Anaphora: The case of Japanese*. Ph.D. thesis, Stanford University, Stanford, CA.

Yeun-Bae Kim and Terumasa Ehara. 1995. Zero-subject resolution method based on probabilistic inference with evaluation function. In *Proceedings of the 3rd Natural Language Processing Pacific- Rim Symposium*, pages 721—727.

YOUNG-JOO Kim. 2000. Subject/object drop in the acquisition of korean: A cross-linguistic comparison. In *Journal of East Asian Linguistics 9.4*, pages 325–351.

Nikita Kitaev, Steven Cao, and Dan Klein. 2018. Multilingual constituency parsing with self-attention and pre-training. *arXiv preprint arXiv:1812.11760*.

Fang Kong and Guodong Zhou. 2010. A tree kernel-based unified framework for chinese zero anaphora resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics*, pages 882–891.

Fang Kong, Min Zhang, and Guodong Zhou. 2019. Chinese zero pronoun resolution: A chain-to-chain approach. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(1):1–21.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of bert. In *arXiv preprint arXiv:1908.08593*.

Charles N. Li and Sandra A. Thompson. 1979. Third person pronouns and zero anaphora in chinese discourse. In *Syntax and Semantics*, volume 12: Discourse and Syntax, pages 311–335. Academic Press.

Ting Liu, Yiming Cui, Qingyu Yin, Weinan Zhang, Shijin Wang, and Guoping Hu. 2017. Generating and exploiting large-scale pseudo training data for zero pronoun resolution. In *arXiv preprint arXiv:1606.01603*.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. The penn arabic treebank: Building a large-scale annotated arabic corpus. In *NEMLAR conference on Arabic language resources and tools*, volume 27, pages 466–467. Cairo.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank.

Claudiu Mihăilă, Iustina Ilisei, , and Diana Inkpen. 2011. Zero pronominal anaphora resolution for the romanian language. In *Research Journal on Computer Science and Computer Engineering with Applications, POLIBITS, 42*.

Ruslan Mitkov and Paul Schmidt. 1998. On the complexity of pronominal anaphora resolution in machine translation. *STUDIES IN FUNCTIONAL AND STRUCTURAL LINGUISTICS*, pages 207–222.

Ruslan Mitkov, Richard Evans, Constantin Orasan, Catalina Barbu, Lisa Jones, and Violeta Sotirova. 2000. Coreference and anaphora: developing annotating tools, annotated resources and annotation strategies. In *Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC2000)*, pages 49–58. Citeseer.

Vinod Nair and Geoffrey Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

Matthew Peters, Sebastian Ruder, and Noah Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. In *arXiv preprint arXiv:1903.05987*.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL-Shared Task. Association for Computational Linguistics, Association for Computational Linguistics.*, pages 1–40.

Luz Rello and Iustina Ilisei. 2009. A rule-based approach to the identification of spanish zero pronouns. In *Proceedings of the Student Research Workshop*, pages 60–65.

Luz Rello, Gabriela Ferraro, and Iria Gayo. 2012. A first approach to the automatic detection of zero subjects and impersonal constructions in portuguese. *Procesamiento del lenguaje natural*, 49:163–170.

Ryohei Sasano and Sadao Kurohashi. 2011. discriminative approach to japanese zero anaphora resolution with large-scale lexicalized case frames. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 758–766.

Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. 2008. A fully-lexicalized probabilistic model for japanese zero anaphora resolution. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 769–776.

Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. 2009. The effect of corpus size on case frame acquisition for discourse analysis. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 521–529.

Kazuhiro Seki, Atsushi Fujii, and Tetsuya Ishikawa. 2002. A probabilistic method for analyzing japanese anaphora integrating zero pronoun detection and resolution. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, pages 1–7.

Sho Shimazu, Sho Takase, Toshiaki Nakazawa, and Naoaki Okazaki. 2020. Evaluation dataset for zero pronoun in japanese to english translation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3630–3634.

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

Lydia White. 1985. The "pro-drop" parameter in adult second language acquisition. *Language learning*, 35(1):47–61.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. In *arXiv preprint arXiv:1609.08144*.

Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural language engineering*, 11(2):207.

Souta Yamashiro, Hitoshi Nishikawa, and Takenobu Tokunaga. 2018. Neural japanese zero anaphora resolution using smoothed large-scale case frames with word embedding. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*.

Ching-Long Yeh and Yi-Chun Chen. 2006. Zero anaphora resolution in chinese with shallow parsing. In *Journal of Chinese Language and Computing 17 (1)*, pages 41–56.

Qingyu Yin, Yu Zhang, Weinan Zhang, and Ting Liu. 2016. A deep neural network for chinese zero pronoun resolution. In *arXiv preprint arXiv:1604.05800.*

Qingyu Yin, Yu Zhang, Weinan Zhang, and Ting Liu. 2017. Chinese zero pronoun resolution with deep memory network. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1309–1318.

Qingyu Yin, Yu Zhang, Weinan Zhang, Ting Liu, and William Yang Wang. 2018. Zero pronoun resolution with attention-based neural network. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 13–23.

Qingyu Yin, Weinan Zhang, Yu Zhang, and Ting Liu. 2019. Chinese zero pronoun resolution: A collaborative filtering-based approach. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(1):1–20.

Katsumasa Yoshikawa, Masayuki Asahara, and Yuji Matsumoto. 2011. Jointly extracting japanese predicate-argument relation with markov logic. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1125–1133.

Kei Yoshimoto. 1988. Identifying zero pronouns in japanese dialogue. In *Coling Budapest 1988 Volume 2: International Conference on Computational Linguistics*.

Koichiro Yoshino, Shinsuke Mori, and Tatsuya Kawahara. 2013. Predicate argument structure analysis using partially annotated corpora. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 957–961.

Shanheng Zhao and Hwee Tou Ng. 2007. Identification and resolution of chinese zero pronouns: A machine learning approach. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 541–550.