ZPR^2: Joint Zero Pronoun Recovery and Resolution using Multi-Task Learning and BERT

Linfeng Song[1], Kun Xu[1], Yue Zhang[2,3], Jianshu Chen[1] and **Dong Yu[1]**
[1]Tencent AI Lab, Bellevue, WA, USA
[2]School of Engineering, Westlake University, China
[3]Institute of Advanced Technology, Westlake Institute for Advanced Study, China

Abstract

Zero pronoun recovery and resolution aim at recovering the dropped pronoun and pointing out its anaphoric mentions, respectively. We propose to better explore their interaction by solving both tasks together, while the previous work treats them separately. For zero pronoun resolution, we study this task in a more realistic setting, where no parsing trees or only automatic trees are available, while most previous work assumes gold trees. Experiments on two benchmarks show that joint modeling significantly outperforms our baseline that already beats the previous state of the arts. Our code is available at https://github.com/freesunshine0316/lab-zp-joint.

1 Introduction

Zero pronoun (ZP) is a linguistic phenomenon where a pronoun is dropped for simplicity. Figure 1 shows an example, where two pronouns at positions ϕ_1 and ϕ_2 are omitted. They both refer to "警方 (The police)" in the sentence beginning and their original form is "他们 (they)". The situation of dropping pronouns happens in most languages. While this phenomenon is not frequent in non-pro-drop languages, such as English, it is extremely severe for pro-drop languages, such as Chinese. In addition, dropped pronouns happens more frequently in conversations than in news. Our preliminary statistics of Chinese shows that 59.2% pronouns are dropped in a corpus of casual dialogues domain, while the number is just 41.6% in another data of broadcast news.

In NLP, dropped pronouns can cause loss of important information, such as the subject or object of the central predicate in a sentence, introducing ambiguity to applications such as machine translation (Nakaiwa and Shirai, 1996; Wang et al., 2016; Takeno et al., 2016), question answering (Choi et al., 2018; Reddy et al., 2019; Sun et al., 2019; Chen and Choi, 2016) and dialogue understanding (Chen et al., 2017; Rolih, 2018). As a result, zero pronouns have recently received much research attention (Liu et al., 2017; Yin et al., 2018a,b). We study Chinese zero pronoun in dialogue settings.

[警方] 怀疑 这 是 一起 黑枪 案件 , ϕ_1 将 枪械 和 皮包 交送 市里 ϕ_2 以 清理 案情 。

[The police] suspected that this is a criminal case about illegal guns , ϕ_1 brought the guns and bags to the city ϕ_2 to deal with the case .

Figure 1: An zero pronoun example and its English translation, where ϕ_1 and ϕ_2 are zero pronouns pointing to the span in square brackets.

There are two long-existing tasks namely *zero pronoun recovery*, which aims at recovering the original pronoun (such as "他 (he)" and "她 (she)"), and *zero pronoun resolution*, where the goal is to pinpoint the mention that each dropped pronoun refers to. Intuitively, the results of the two tasks highly interact with each other. Taking Figure 1 as an example, it will be much easier to resolute ϕ_1 to "警方 (The police)" rather than "黑枪 案件 (criminal case about illegal guns)" if we know ϕ_1 corresponds to "他们 (they)". Similarly, it would be more likely to recover ϕ_1 as "他们 (they)" than other candidate pronouns, if we know ϕ_1 points to "警方 (The police)".

Despite their high correlation, previous work considers them as irrelevant tasks, solving them separately by different models. This can waste training resources, as each task has a limited number of labeled instances, and thus data sparsity can limit model performance. Besides, we believe that it is unnecessary to keep a specific model for each task, as they can be close enough to be solved together. In addition, most *zero pronoun resolution* research (Chen and Ng, 2013, 2016; Kong and Zhou, 2010; Iida and Poesio, 2011; Sasano et al., 2008; Yin et al., 2018b; Yang et al., 2019) as-

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5429–5434
July 5 - 10, 2020.

sumes gold trees being available with the positions of zero pronouns, which is unrealistic in practical applications. During decoding, a zero pronoun resolution model has to rely on automatic trees and zero pronoun detection, thus suffering from error propagation.

In this paper, we propose to jointly solve both tasks under a heterogeneous multi-task learning framework, where each data point only has the annotation of one task, to benefit from the supervised data of both tasks. As the result, we enjoy the benefit of more supervised training data. To improve the robustness of heterogeneous training and introduce more supervision, we introduce *zero pronoun detection*, a common sub-task for both ZP resolution and recovery. Zero pronoun detection is a binary-classification task aiming to detect whether a word space has a dropped pronoun.

We consider ZP recovery as a sequence labeling task, regarding whether each word space has a dropped pronoun and what type the pronoun is. ZP resolution is solved as extractive reading comprehension (Rajpurkar et al., 2016), where each word space is taken as a query and its anaphoric mentions are treated as the answers. For non-ZP spaces where there is no corresponding anaphoric mentions, we assign the sentence beginning (span [0,0]) as the answer.

Experiments on two benchmarks, OntoNotes 5.0[1] (ZP resolution) and BaiduZhdiao (Zhang et al., 2016) (ZP recovery), show that joint modeling gives us 1.5+ absolute F1-score gains for both tasks over our very strong baselines using BERT (Devlin et al., 2019). Our overall system gives an dramatic improvement of 3.5 F1 points over previous state-of-the-art results on both tasks.

2 Related work

Previous work considers zero pronoun resolution and recovery separately. For zero pronoun recovery, existing methods can be classified according to the types of annotations they use. One line of work (Yang et al., 2015, 2019) simply relies on the human annotations, solving the task as sequence labeling. The other line of work (Chung and Gildea, 2010; Xiang et al., 2013; Wang et al., 2016) mines weak supervision signals from a large bilingual parallel corpus, where the other language is non-pro-drop with fewer pronoun drops. The latter requires massive training data, and the MT performance is the primary goal, thus we follow the first line of research using human-annotated data.

Rao et al. (2015) studied zero pronoun resolution in multi-turn dialogues, claiming that their model does not rely on parsing trees to extract ZP positions and noun phrase as resolution candidates. However, they only consider the dropped pronouns that correspond to one of the dialogue participant. As a result, they only explore a small subset of the entire ZP resolution problem, and their task is closer to zero pronoun recovery. Most similar to our work, Liu et al. (2017) converted zero pronoun resolution as a machine reading comprehension task (Rajpurkar et al., 2016) in order to automatically construct a large-scale pseudo dataset for model pretraining. However, their model finetuning and evaluation with benchmark data still rely on human-annotated trees and gold zero pronoun positions. As a result, it is still uncertain what performance a model can achieve without such gold inputs. We address both issues in the joint task.

Our work is inspired by the recent advances of heterogeneous multi-task learning using BERT (Devlin et al., 2019), which combines the supervised data of several related tasks to achieve further improvements. In particular, Liu et al. (2019) utilize this framework to jointly solve GLUE tasks (Wang et al., 2019). But their experiments show that multi-task learning does not help across all tasks. Our work takes a similar spirit, and our contribution is mainly on the zero pronoun tasks. In addition, we find that it helps the robustness of multi-task learning to add a common sub-task (e.g. zero pronoun detection in our case) for additional supervision and alleviating annotation variances, if such a sub-task is available.

3 Model

As shown in Figure 2, we model ZP recovery (f_{rec}), ZP resolution (f_{res}), and the auxiliary ZP detection (f_{det}) task with multi-task learning, where BERT (Devlin et al., 2019) is used to represent each input sentence $s_1 \dots s_N$ of N words to provide shared features.

3.1 Zero pronoun recovery

ZP recovery is to restore any dropped pronouns for an input text. Since pronouns are enumerable (e.g. there are 10 types for Chinese), we cast this task into a classification problem for each word space. Taking some shared input representations

[1] https://catalog.ldc.upenn.edu/LDC2013T19

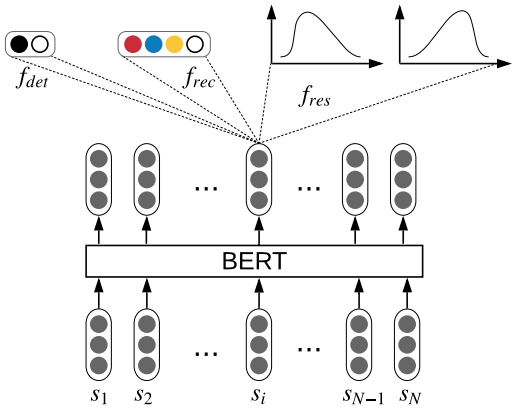

Figure 2: Model framework.

$\boldsymbol{h}_0, \boldsymbol{h}_1, \ldots, \boldsymbol{h}_N$, the probability for recovering pronoun p_i at the space between s_{i-1} and s_i is:

$$p(p_i|X, i) = \text{softmax}(\boldsymbol{W}_r \boldsymbol{h}_i + \boldsymbol{b}_r) \quad (1)$$

where $\boldsymbol{W}_r$ and $\boldsymbol{b}_r$ are model parameters.

3.2 Zero pronoun resolution

Our zero pronoun resolution task is to predict the span that each dropped pronoun points to, while the gold ZP positions are not available. One potential solution is executing zero pronoun recovery first and utilize that information, while this introduces error propagation. Conversely, we manually assign span "(0,0)" for non-ZP positions. This will not introduce conflicts, as position "0" corresponds to the special token `[CLS]` for BERT encoding and thus no real spans can be "(0,0)".

We cast the resolution task for each word space (such as between s_{i-1} and s_i) as machine reading comprehension (MRC) (Rajpurkar et al., 2016), where a resolution span corresponds to a MRC target answer. Following previous work on MRC, we separately model the start (r_i^{st}) and end (r_i^{ed}) positions for each span with self-attention:

$$\begin{aligned} p(r_i^{st}|X, i) &= \text{SelfAttn}_{st}(\boldsymbol{H}, \boldsymbol{h}_i) \\ p(r_i^{ed}|X, i) &= \text{SelfAttn}_{ed}(\boldsymbol{H}, \boldsymbol{h}_i) \end{aligned} \quad (2)$$

where $\boldsymbol{H} = [\boldsymbol{h}_0, \ldots, \boldsymbol{h}_N]$ is the concatenation of all word states, and $\text{SelfAttn}_{st}()$ and $\text{SelfAttn}_{ed}()$ are the self-attention modules for predicting the start and end positions of each ZP resolution span. The probability for the whole span r_i is:

$$p(r_i|X, i) = p(r_i^{st}|X, i)p(r_i^{ed}|X, i) \quad (3)$$

3.3 Auxiliary task: zero pronoun detection

We also introduce pronoun detection as an auxiliary task to enhance multi-task training. This task is to determine whether each word space has a dropped pronoun. Similar with zero pronoun recovery, we formulate it as binary classification:

$$p(d_i|X, i) = \text{softmax}(\boldsymbol{W}_d \boldsymbol{h}_i + \boldsymbol{b}_d) \quad (4)$$

where d_i is the binary detection result. $\boldsymbol{W}_d$ and $\boldsymbol{b}_d$ are model parameters.

3.4 Encoding input with BERT

Given an input sentence $s_1, \ldots, s_N$, we use BERT to encode them into a sequence of input features shared across all our tasks. We append the `[CLS]` token to inputs, before sending them to BERT. Our task features are represented as $\boldsymbol{h}_0, \boldsymbol{h}_1, \ldots, \boldsymbol{h}_N$, where $\boldsymbol{h}_0$ corresponds to token `[CLS]`.

3.5 Training

We train our model on the combined and shuffled data of both tasks to leverage more supervision signals. Each data instance only contains the annotation of either ZP recovery or resolution, thus the loss for one example is defined as:

$$\begin{aligned} loss = -\sum_{i \in 1..N} \Big(& \alpha \log p(\mathrm{p}_i|X, i) \\ & - \beta \log p(\mathrm{r}_i|X, i) - \gamma \log p(\mathrm{d}_i|X, i) \Big) \end{aligned} \quad (5)$$

where α, β and γ are the coefficients for the tasks. For α and β, the value of is 1 if the corresponding supervision exists, otherwise it is 0. We empirically set the value of γ to 0.1, as the supervision of ZP detection exists for all instances, and we do not want this auxiliary loss signal to be too strong.

4 Experiments

We study the effectiveness of jointly modeling ZP resolution, recovery and detection.

4.1 Data and setting

We take two benchmark datasets: BaiduZhidao (Zhang et al., 2016), a benchmark for ZP recovery, and OntoNotes 5.0, a benchmark for ZP resolution. For BaiduZhidao, we use the version cleaned by Yang et al. (2019), containing 5504, 1175 and 1178 instances for training, development and testing, respectively. OntoNotes 5.0 has 36487 training and 6083 testing instances, and we separate 20% training instances for development.

Model	OntoNotes 5.0 (RES)			BaiduZhidao (REC)			Avg. F1
	P	R	F	P	R	F	
ZPMN (Yin et al., 2017)	18.5	**29.3**	22.7	–	–	–	–
NDPR-W (Yang et al., 2019)	–	–	–	38.60	**50.12**	43.36	–
BERT	26.87	22.43	24.45	43.50	47.30	45.32	34.89
BERT-MTL	24.55	25.49	25.01	41.63	48.22	44.68	34.85
BERT-MTL w/ detection	**30.96**	22.51	**26.07**	**46.09**	47.54	**46.81**	**36.44**

Table 1: Main results for ZP resolution and recovery, where RES and REC are short for resolution and recovery.

Model	P	R	F
Gold Tree + Gold ZP			
ZPMN (Yin et al., 2017)	55.1	54.8	54.9
AttentionZP (Yin et al., 2018b)	–	–	57.3
Our model	**59.40**	**57.61**	**58.49**
Gold Tree + Auto ZP			
ZPMN (Yin et al., 2017)	31.1	**39.4**	34.8
Our model	**42.56**	32.03	**36.55**

Table 2: ZP resolution with gold trees and ZP positions.

Method	***Auto Tree + Auto ZP***		
	P	R	F
Our model	30.96	22.51	26.07
w/ auto tree cons.	36.13	32.32	34.12

Table 3: Resolution using automatic trees as constraint.

We choose the official pretrained Chinese BERT-base model[2]. Models are trained with Adam (Kingma and Ba, 2014) with a learning rate of 10^{-5} and a warm-up proportion of 10%. To avoid overfitting, we apply $l2$ norm for BERT parameters with a coefficient of 0.01. Models are selected by early stopping with development results.

4.2 Main results

Table 1 shows the results for both resolution and recovery tasks, where *ZPMN* and *NDPR-W* show the state-of-the-art performances without relying on any gold syntactic information. *ZPMN* treats zero pronoun resolution as a classification task over noun phrase candidates, and the final result is selected using an attention mechanism. *NDPR-W* studies zero pronoun recovery in dialogues by modeling all dialogue history.

For our models, *BERT* represents finetuning BERT only on one task, *BERT-MTL* means jointly finetuning BERT on both tasks with multi-task learning (as shown in Figure 2), and *BERT-MTL w/ detection* is our model with auxiliary detection loss. Using BERT already gives us much better performances than the previous state-of-the-art results. Initial usage of heterogeneous multi-task learning helps ZP resolution, while hurting ZP recovery, and one potential reason is that the ZP resolution dataset (OntoNotes 5.0) has much more instances than the ZP recovery dataset (BaiduZhidao). This problem is alleviated by introducing the auxiliary ZP detection task due to the following possible reasons. Most importantly, ZP detection is very close to ZP recovery (binary vs multi-class), thus this extra supervision helps to alleviate the data magnitude imbalance problem. Besides, ZP detection introduces more useful training signals to the overall training process.

4.3 More analysis on ZP resolution

We also evaluate on other previously studied settings, where gold trees or even gold ZP positions are given. As *ZPMN* also reported strong performances cross these settings, we take this model as a baseline for comparison.

Using gold trees and ZP positions Since most previous work on ZP resolution uses gold syntactic trees and/or ZP positions, we also investigate our performance under these settings. In particular, we take the noun phrases and/or ZP positions from gold trees to serve as constraints. Besides, our model is only trained on the ZP positions when they are given. Table 2 shows the results, *AttentionZP* gives the previous state-of-the-art performance under the *Gold Tree + Gold ZP* setting. Our model outperforms *AttentionZP* by a significant margin. Beside, we also report the best performance, which significantly outperforms the previous best system (*ZPMN*) under the *Gold Tree + Auto ZP* setting, where only gold trees are available.

Effectiveness of automatic trees Currently, our model considers all free spans when making a resolution decision. Using automatic tree can greatly limit the search space, while that could introduce errors. We conduct a preliminary comparison as shown in Table 3, where such a constraint dramatically helps the performance. But, this is based on the assumption that the target-domain syntactic parsing is very accurate, as our ZP resolution data (OntoNotes 5.0) is mostly collected from

[2]https://github.com/google-research/bert

broadcast news. The F1 score using automatic trees (34.12) is close to the score using gold trees (36.55 in Table 2), which also indicates the conjecture above. As a result, we may expect a performance drop for web and biomedical domains, where the parsing accuracies are much lower.

5 Conclusion

We studied the effectiveness of jointly modeling ZP recovery and resolution using the recently introduced *multi-task learning + BERT* framework. To alleviate the data magnitude imbalance problem, we introduce ZP detection as a common auxiliary sub-task for extra supervision. Experiments on two benchmarks show that our model is consistently better than previous results under various settings, and that the auxiliary ZP detection sub-task can make the training process more robust.

References

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

Henry Y Chen, Ethan Zhou, and Jinho D Choi. 2017. Robust coreference resolution and entity linking on dialogues: Character identification on tv show transcripts. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 216–225.

Yu-Hsin Chen and Jinho D Choi. 2016. Character identification on multiparty conversation: Identifying mentions of characters in tv shows. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 90–100.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184.

Tagyoung Chung and Daniel Gildea. 2010. Effects of empty categories on machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 636–645. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

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.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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*, pages 882–891.

Ting Liu, Yiming Cui, Qingyu Yin, Wei-Nan Zhang, Shijin Wang, and Guoping Hu. 2017. Generating and exploiting large-scale pseudo training data for zero pronoun resolution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 102–111.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 4487–4496.

Hiromi Nakaiwa and Satoshi Shirai. 1996. Anaphora resolution of japanese zero pronouns with deictic reference. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, pages 812–817. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Sudha Rao, Allyson Ettinger, Hal Daumé III, and Philip Resnik. 2015. Dialogue focus tracking for zero pronoun resolution. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 494–503.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Gabi Rolih. 2018. Applying coreference resolution for usage in dialog systems.

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-Volume 1*, pages 769–776. Association for Computational Linguistics.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Shunsuke Takeno, Masaaki Nagata, and Kazuhide Yamamoto. 2016. Integrating empty category detection into preordering machine translation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 157–165.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. *the Proceedings of ICLR*.

Longyue Wang, Zhaopeng Tu, Xiaojun Zhang, Hang Li, Andy Way, and Qun Liu. 2016. A novel approach to dropped pronoun translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 983–993.

Bing Xiang, Xiaoqiang Luo, and Bowen Zhou. 2013. Enlisting the ghost: Modeling empty categories for machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 822–831.

Jingxuan Yang, Jianzhuo Tong, Si Li, Sheng Gao, Jun Guo, and Nianwen Xue. 2019. Recovering dropped pronouns in chinese conversations via modeling their referents. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 892–901.

Yaqin Yang, Yalin Liu, and Nianwen Xue. 2015. Recovering dropped pronouns from chinese text messages. 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 309–313.

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. 2018a. Deep reinforcement learning for chinese zero pronoun resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 569–578.

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

Wei-Nan Zhang, Ting Liu, Qingyu Yin, and Yu Zhang. 2016. Neural recovery machine for chinese dropped pronoun. *arXiv preprint arXiv:1605.02134*.