Modelling Pro-drop with the Rational Speech Acts Model

Guanyi Chen[1], Kees van Deemter[12], Chenghua Lin[2]
[1]Department of Information and Computing Sciences, Utrecht University
[2]Department of Computing Science, University of Aberdeen
{g.chen, c.j.vandeemter}@uu.nl, chenghua.lin@abdn.ac.uk

Abstract

We extend the classic Referring Expressions Generation task by considering zero pronouns in “pro-drop” languages such as Chinese, modelling their use by means of the Bayesian Rational Speech Acts model (Frank and Goodman, 2012). By assuming that highly salient referents are most likely to be referred to by zero pronouns (i.e., pro-drop is more likely for salient referents than the less salient ones), the model offers an attractive explanation of a phenomenon not previously addressed probabilistically.

1 Introduction

Languages such as Chinese and Japanese make liberal use of zero pronouns (ZP) (Huang, 1984). The analysis of Wang et al. (2018) on a large Chinese-English parallel dialogue corpus shows that 26% of the English pronouns are dropped in Chinese. Such an abundant use of zero pronouns has been a key factor in linguist’s idea (Huang, 1984, 1989) that Chinese is a *“cool”* language or a *discourse-oriented* language (Cao, 1979), i.e., one that relies heavily on context.

To exemplify zero pronouns in Chinese, consider the question “你今天看见比尔了吗?” (*Did you see Bill today?*). A Chinese speaker can respond in a variety of shorter expressions which are equivalent to “我看见他了” (*Yes, I saw him*), for example, “∅看见他了” (*Yes, ∅ saw him*), “我看见∅了” (*Yes, I saw ∅*), or even “∅看见∅了” (*Yes, ∅ saw ∅*). Here the ∅ symbol indicates the place from where a pronoun appears to have been “dropped” from a full sentence.

Generating zero pronouns (only) where they are appropriate is a difficult challenge for Referring Expression Generation (REG) (Van Deemter, 2016), and more specifically for the task of choosing *referential form*, a key step in the classic Natural Language Generation (NLG) architecture (Reiter and Dale, 2000). Traditionally, choosing referential form is framed as modelling speakers’ behaviour of deciding whether entities are referred to using a pronoun, a proper name, or a description. However, for “cool” languages, an extra option, namely of choosing a zero pronoun, needs to be added (Yeh and Mellish, 1997) for fully simulating speakers’ behaviour.

In this paper, we model the use of zero pronouns in Chinese with the Rational Speech Acts (RSA) model (Frank and Goodman, 2012) by assuming that speakers tend to choose a ZP if it is salient enough for successful communication (see §2). For computing discourse salience, we focus on ZPs that are *recoverable*, meaning that they either refer anaphorically to an entity mentioned earlier in the text (i.e., anaphoric ZPs, or AZPs for short), or to the speaker or hearer (i.e., deictic non-anaphoric ZPs or DNZPs for short) (Zhao and Ng, 2007); a ZP is *unrecoverable* if it cannot be linked to any referent, for example:

(1) ∅有 二十三 项 高新技术
∅ has 23 CLASSFIER high-tech
项目 进区 开发
projects in.the.zone under.development
‘there are 23 high-tech projects under development in the zone’

in which the ∅ cannot be recovered.

2 Related Work

Pro-drop raises challenges for a number of NLP tasks including, machine translation (MT), co-reference resolution, and REG. When translating from a pro-drop language, recovering the dropped pronouns of the source language can improve the overall performance of MT (Wang et al., 2016,

Proceedings of The 11th International Natural Language Generation Conference, pages 159–164,
Tilburg, The Netherlands, November 5-8, 2018. ©2018 Association for Computational Linguistics

2018). Co-reference resolution of ZPs has been widely explored with a variety of techniques including the centring theory (Rao et al., 2015), statistical machine learning (Zhao and Ng, 2007; Chen and Ng, 2014, 2015), deep learning (Chen and Ng, 2016; Yin et al., 2016, 2017) and reinforcement learning (Yin et al., 2018). REG of ZPs for "cool" languages has been addressed through rule-based methods (Yeh and Mellish, 1997) including centring theory (Yamura-Takei et al., 2001) (for Japanese), but we are not aware of any testable computational account.[1] We offer such an account, along probabilistic lines.

Some discourse theories suggest that speakers choose referring expressions (REs) by considering discourse salience (Givón, 1983), i.e., speakers tend to choose pronouns if they believe the referent is highly salient. The intuition behind is that a highly salient referent tends to be highly prominent in the mind of the speaker and/or hearer. Orita et al. (2015) shared a similar view and argued that highly salient REs are highly *predictable*, so they are referred with pronouns (as opposed to full NPs) more often than the less salient ones.

A theory that is sometimes used for explaining the relation between discourse salience and human choice of referential forms is Uniform Information Density (UID) (Jaeger and Levy, 2007). UID asserts that speaker tends to optimise information density (quantity of information) of the utterances to achieve optimal communication. In other words, speakers tend to drop a RE when the referent of the RE is predictable (or recoverable), and vise versa.

Apart from salience, production cost (Rohde et al., 2012) and the listener models (Bard et al., 2004), meaning the models that how speakers model listeners' interpretation of the utterance, also have impact on language production. It suggests to us that the salience of the referent may not be enough for modelling speakers' choice. The RSA model (see §3) used in this paper is possible to take all these factors into consideration.

3 Methodology

3.1 The Rational Speech Acts Model

The Rational Speech Acts (RSA) model (Frank and Goodman, 2012) has been used for a variety of tasks including modelling speakers' referential choice between pronouns and proper names (Orita et al., 2015), the selection of attributes for referring expressions (Monroe and Potts, 2015), and the generation of colour references (Monroe et al., 2017, 2018). The key idea of RSA is to model human communication by assuming that a rational listener P_L uses Bayesian inference to recover a speaker's intended referent r_s for word w under context C. In this way, RSA claims to offer not only accurate models, but highly explanatory ones as well. Formally, P_L is defined as

$$P_L(r_s|w, C) = \frac{P_S(w|r_s, C)P(r_s)}{\sum_{r' \in C} P_S(w|r', C)P(r')}, \quad (1)$$

where r' denotes a referent in context C, $P(r_s)$ represents the discourse salience of r_s, P_S is the speaker model defined by an exponential utility function:

$$P_S(w|r_s, C) = e^{\alpha(I(w;r_s,C)-C(w))}. \quad (2)$$

Here $I(w; r_s, C)$ is the informativeness of word w, $C(w)$ represents the speech cost.

Orita et al. (2015) extended the RSA by assuming that speakers estimate listener's interpretation of the (form of) RE w based on discourse information. The speaker chooses w by maximising the listener's belief in the speaker's intended referent r_s in relation to the speaker's speech cost $C(w)$, where the cost is estimated according to the complexity of the utterance, such as the length of w:

$$\begin{aligned} P_S(w|r_s) &\propto P_L(r_s|w) \cdot \frac{1}{C(w)} \\ &= \frac{P(w|r_s, C)P(r_s)}{\sum_{r'} P(w|r', C)P(r')} \cdot \frac{1}{C(w)} \end{aligned} \quad (3)$$

Here $P_L(r_s|w)$ estimates the informativeness of w, and $P(w|r_s, C)$ estimates the likelihood (according to the speaker) that the listener guesses that the speaker used w to refer to r_s.

3.2 Modelling Pro-drop with the RSA Model

We model the decision of whether to use a ZP-based on the formulation expressed in Eq. 3. The speaker model is $P_S(z|r_s)$, which is the probability that the speaker uses ZP (i.e., drops the RE). We assume that the speaker makes a binary choice (i.e., $z = \{1, 0\}$), with $z = 1$ indicating a ZP and $z = 0$ indicating a non-zero form of RE (NZRE). Note that whether the speaker uses a pronoun or

[1] E.g., Yeh and Mellish (1997) did not offer a precise definition of some of the syntactic constraints and the notion of salience that they were using.

a proper name is not in the scope of this model. To simulate the speaker's choice, we need to estimate the dropping probability $P(z|r_s)$, the discourse salience of the referent $P(r_s)$, and the cost $C(z)$.

According to the UID theory (see §2), if a RE is recoverable, then the speaker prefers a ZP over a NZRE to maximise the information density since a ZP is shorter than any other referential form. In that sense, we follow Orita et al. (2015) to estimate the **cost function** $C(z)$ based on the length of the RE, i.e., the total number of words the RE contains. However, the length of the NZRE is not known in advance, thus we use the average length of a set of REs W instead:

$$C(z = 0) = \text{average\_length}(W) + 1 \tag{4}$$

We experimented with two ways of calculating the average length: (i) *global average length*, meaning that W is the set of all referring expressions in the corpus, and (ii) *local average length*, in which W is the set of expressions that can refer to referent r_s. For instance, if r_s is "*Barack Obama*", then given a corpus for computing local average length in which *he* is referred to, W might be the set {*Barack Obama*, *Obama*, *he*, *former president*}. The cost of a zero pronoun is always $C(z = 1) = 1$, which means no discount on $P(z = 1|w)$ and the plus 1 in Eq. 4 is to make the cost of choosing NZRE different from choosing ZP if W only contains pronouns (i.e., if length equals to 1).

We assume that the **dropping probability** $P(z|r_s)$ is dependent on whether the referent r_s is one of the participants in the dialogue (i.e., speaker or listener). For example, in the OntoNote 5.0 corpus, 30% of maximally salient entities are dropped, which is much higher than the 10% dropping rate of non-maximally salient entities. If r_s is one of the participants, we call it *maximally salient entity* (denoted as s). Otherwise, r_s is called *non-maximally salient entity* (denoted as ns). This assumption causes AZP and DNZP to have different proportions in the predicted results. Suppose $P(z = 1|r_s = \text{ns}) = a$ and $P(z = 1|r_s = \text{s}) = b$, then we have $a < b$, which implies that the speaker thinks the listener expects a maximally salient entity (i.e., speaker or listener).

Let $\alpha = \frac{a}{b}$ be the *dropping ratio*, then the probability of dropping a noun phrase that refers to the speaker is:

$$\begin{aligned} P_S(\text{ZP}|\text{Speaker}) &\propto P_L(\text{Speaker}|\text{ZP}) \cdot \frac{1}{C(z=1)} \\ &= \frac{P(\text{ZP}|\text{Speaker})P(\text{Speaker})}{\sum_{r'} P(\text{ZP}|r')P(r')} \cdot \frac{1}{C(z=1)} \\ &= \frac{N_{\text{Speaker}}}{\alpha \cdot N_{\text{NS}} + N_{\text{S}}} \cdot \frac{1}{C(z=1)} \end{aligned} \tag{5}$$

$P(\text{Speaker})$ is the **salience** of the speaker.[2] In general, we take the salience of a referent x to be in proportion to N_x, which is the number of times that x has been referred to in the preceding discourse, hence the use of N_{Speaker}, N_{S}, and N_{NS} in the equation. Note that $N_{\text{S}} + N_{\text{NS}}$ is the total number of REs in the preceding discourse.

Equation 5 shows that modelling the dropping probability for maximally salient entities and non-maximally salient entities differently acts as a discount for the number of referents that the ZP can refer to when predicting DNZP. Similarly, using the dropping ratio α, the dropping probability for AZPs is estimated as:

$$P_S(\text{AZP}|\text{Speaker}) = \frac{N_{\text{AZP}}}{N_{\text{NS}} + \frac{1}{\alpha} N_{\text{S}}} \tag{6}$$

which can be seen as adding a penalty.

The frequencies counted above are all based on the whole preceding discourse of a referent, which might not be reasonable for predicting ZPs. We hypothesise that the informativeness of a ZP depends on only a part of the preceding context. We tested two possible set-ups. One is setting a discourse window to limit the number of sentences that the simulator can look back to. The other uses recency (Chafe, 1994). Following Orita et al. (2015), we replace each count with: $Count(r_i, r_j) = e^{-d(r_i, r_j)/a}$, where r_j is the same referent as the r_i that has previously been referred to and d is the number of sentences between two REs. Instead of taking the direct raw count 1, $Count(r_i, r_j)$ decays exponentially with respect to how far it is from the predicting RE. The RE that has larger distance contributes less to the overall count of that referent.

For NZREs ($z = 0$), we assume that the number of times that the referent has been referred to is equal to the total number of referents referred to by that NZRE. Thus, the speaker believes that the listener can always resolve the reference by giving

[2] Our use of the term salience is similar to Hovy et al. (2006)'s use of "recoverability".

them a NZRE. In other words, their informativeness equals 1.

4 Experiments

4.1 The Dataset

We tested our model on the Chinese portion of OntoNotes Release 5.0 data[3] (Hovy et al., 2006), which has been widely used in (ZP) co-reference resolution tasks. The corpus contains 1,729 documents, including 143620 referring expressions. In Table 1, there is the basic statistics about the recoverable zero pronouns in OntoNotes corpus.

# of Recoverable Zero Pronouns	17,129
# of Anaphoric ZPs	14,675
# of Deictic Non-anaphoric ZPs	2,454

Table 1: Basic statistics of different types of recoverable ZPs in OntoNotes

Baseline. In this work, we used the modified rule 1 in Yeh and Mellish (1997), i.e., the RE in the subject position will be a ZP if it was referred to in the immediately preceding sentence, as the baseline. The modification is inspired by the fact that 99.2% ZPs in OntoNotes corpus are in the subject position.

4.2 Experiment Results

Table 2 shows the results (reported in accuracy) of various models on the OntoNote dataset. The dropping ratio α was empirically set to 0.1 and the decay parameter a of recency was set to 0.8. The window size was 1, so the simulator only looks at the current sentence and preceding sentence.

As expected, the models that look back to the whole preceding discourse perform badly on predicting ZPs (i.e., 8.35% of accuracy), especially DNZPs. They tend to predict all REs as NZREs, which even performs worse than the model using simple rule (i.e., the baseline). In contrast, limiting the discourse history by applying discourse windows or replacing frequency with recency have a negative impact on predicting NZREs, more specifically pronouns. Such an impact is caused by the idea that every NZRE can always be resolved by the listener, which is not correct for pronouns. However, so far, we cannot calculate the informativeness of pronouns properly since we do not know which referent (speaker or listener) a deictic pronoun in the corpus refers to. For example, in the corpus, both the speaker and listener will use “I” to refer to themselves, so we don’t know whether “I” refers to the speaker or the listener. This setting will lead to over-estimation of the informativeness of pronouns. On the other hand, computing cost by average length (as we do) over-estimates the costs of pronouns, whose lengths are generally shorter than proper names.

The baseline model’s performance is not bad, especially for predicting AZPs. This is partly because the rule predicts that all REs in object position are NZREs and this is nearly always correct. (Recall that 99.84% REs in object position are NZREs). At the same time, if the referent was referred to in the immediately preceding sentence (as the baseline model requires), then it is clearly more salient than if it wasn’t. The baseline model is therefore quite similar to the model with discourse window, but its decisions are made in a simpler way (i.e., based on a simple ”if-then” rule).

With respect to overall accuracy for predicting ZPs and NZREs, models with recency perform similarly to those that use a discourse window. However, recency offers better prediction on AZPs. Adding a dropping ratio could significantly improve the performance on predicting DNZPs without decreasing the accuracies of AZPs and NZREs very much (i.e., accuracy increase from 62.02% to 95.35%). For the choice of cost function, we found that using global average length is the best.

5 Conclusion and Future Work

This paper has explored the possibilities of using the RSA model for probabilistic simulation of speakers’ use of ZPs (i.e., pro-drop), and investigated factors that influence speakers’ choice.

Our model performs respectably yet, as mentioned in Section 4, it under-estimates the probability of choosing a pronoun. Solving this problem will require a more fine-grained annotation of the corpus, indicating which person each occurrence of the deictic pronouns ”I” and ”you” refers to. Once this has been done, we also hope to let the generator distinguish between ZP, pronoun, proper name, and full noun phrase.

When speakers are choosing between pronouns and full NPs, sentence position is known to be rel-

[3] https://catalog.ldc.upenn.edu/ldc2013t19

Discourse	Model	Cost	Total Acc.	ZP Acc.	AZP Acc.	DNZP Acc.	NZRE Acc.
-	baseline	-	78.57	40.88	42.90	28.81	83.67
Discourse Window	full	global	77.10	46.16	38.34	92.95	81.29
		local	81.79	22.53	25.50	4.81	89.81
	-dropping ratio	global	77.05	43.77	41.88	55.09	81.56
		local	81.44	23.67	27.09	3.19	89.26
Recency	full	global	75.64	**50.56**	43.08	**95.35**	79.03
		local	80.08	25.36	28.81	4.77	87.49
	-dropping ratio	global	74.04	50.26	**48.29**	62.02	78.04
		local	79.26	27.47	31.63	2.6	86.28
Whole	full	global	86.24	8.35	5.18	27.30	96.79
		local	**86.67**	3.67	4.27	0.08	**97.91**
	-dropping ratio	global	86.13	6.23	6.38	5.33	96.95
		local	86.61	3.84	4.47	0.04	97.81

Table 2: Accuracies of each model, recall that AZP and DNZP are two sub-categories of ZP.

evant. For example, pronouns are less common in object than in subject position Brennan (1995), which somehow dues to the fact that REs in subject position are more salient than in object position. In the OntoNotes corpus, 99.2% of ZPs appear in subject position; in Chinese, empty categories are acceptable in both subject and object (including the topic position (Huang, 1984)), but even there they are most frequent in subject position. The baseline model introduced in this paper has somehow proved that considering positions works in modelling pro-drop. In future we shall explore the way of combining that factor with the RSA for pro-drop model introduced in this paper.

In future, we will investigate alternative ways to estimate informativeness and costs. For example, it would be natural to use a co-reference resolver for calculating informativeness. Furthermore, one could follow on from (Yamura-Takei et al., 2001; Roh and Lee, 2003) by using elements of centring theory (Grosz et al., 1995) in the definition of cost (e.g., giving Rough Shifts a high cost). Alternatively, one could improve the model by adopting a trainable function for estimating both informativeness and costs.

Acknowledgements

We thank Ehud Reiter, people in the Aberdeen CLAN group and anonymous reviewers for their comments for earlier version of this paper. We also thank Qingyu Yin for his help on pre-processing the datasets.

References

Ellen Gurman Bard, Matthew P Aylett, J Trueswell, and M Tanenhaus. 2004. Referential form, word duration, and modeling the listener in spoken dialogue. *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions*, pages 173–191.

Susan E Brennan. 1995. Centering attention in discourse. *Language and Cognitive processes*, 10(2):137–167.

Fengfu Cao. 1979. *A functional study of topic in Chinese: The first step towards discourse analysis*, volume 3. Student Book Co.

Wallace Chafe. 1994. Discourse, consciousness, and time. *Discourse*, 2(1).

Chen Chen and Vincent Ng. 2014. Chinese zero pronoun resolution: An unsupervised approach combining ranking and integer linear programming. In *AAAI*, pages 1622–1628.

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)*, volume 2, pages 320–326.

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)*, volume 1, pages 778–788.

Michael C Frank and Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Talmy Givón. 1983. *Topic continuity in discourse*. John Benjamins Publishing Company.

Barbara J Grosz, Scott Weinstein, and Aravind K Joshi. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, 21(2):203–225.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60. Association for Computational Linguistics.

C-T James Huang. 1984. On the distribution and reference of empty pronouns. *Linguistic inquiry*, pages 531–574.

C-T James Huang. 1989. Pro-drop in chinese: A generalized control theory. In *The null subject parameter*, pages 185–214. Springer.

T Florian Jaeger and Roger P Levy. 2007. Speakers optimize information density through syntactic reduction. In *Advances in neural information processing systems*, pages 849–856.

Will Monroe, Robert XD Hawkins, Noah D Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *arXiv preprint arXiv:1703.10186*.

Will Monroe, Jennifer Hu, Andrew Jong, and Christopher Potts. 2018. Generating bilingual pragmatic color references. *arXiv preprint arXiv:1803.03917*.

Will Monroe and Christopher Potts. 2015. Learning in the rational speech acts model. *arXiv preprint arXiv:1510.06807*.

Naho Orita, Eliana Vornov, Naomi Feldman, and Hal Daumé III. 2015. Why discourse affects speakers' choice of referring expressions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1639–1649.

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.

Ehud Reiter and Robert Dale. 2000. *Building natural language generation systems*. Cambridge university press.

Ji-Eun Roh and Jong-Hyeok Lee. 2003. An empirical study for generating zero pronoun in korean based on cost-based centering model. In *Proceedings of the Australasian Language Technology Workshop 2003*, pages 30–37.

Hannah Rohde, Scott Seyfarth, Brady Clark, Gerhard Jäger, and Stefan Kaufmann. 2012. Communicating with cost-based implicature: A game-theoretic approach to ambiguity. In *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue*, pages 107–116.

Kees Van Deemter. 2016. *Computational models of referring: a study in cognitive science*. MIT Press.

Longyue Wang, Zhaopeng Tu, Shuming Shi, Tong Zhang, Yvette Graham, and Qun Liu. 2018. Translating pro-drop languages with reconstruction models. *arXiv preprint arXiv:1801.03257*.

Longyue Wang, Zhaopeng Tu, Xiaojun Zhang, Hang Li, Andy Way, and Qun Liu. 2016. A novel approach to dropped pronoun translation. *arXiv preprint arXiv:1604.06285*.

Mitsuko Yamura-Takei, Miho Fujiwara, and Teruaki Aizawa. 2001. Centering as an anaphora generation algorithm: A language learning aid perspective. In *NLPRS*, volume 2001, pages 557–562.

Ching-Long Yeh and Chris Mellish. 1997. An empirical study on the generation of anaphora in chinese. *Computational Linguistics*, 23(1):171–190.

Qingyu Yin, Weinan Zhang, Yu Zhang, and Ting Liu. 2016. A deep neural network for chinese zero pronoun resolution. *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. Deep reinforcement learning for chinese zero pronoun resolution. *arXiv preprint arXiv:1806.03711*.

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 (EMNLP-CoNLL)*.