The Users Who Say 'Ni': Audience Identification in Chinese-language Restaurant Reviews

Rob Voigt **Dan Jurafsky**
Stanford University
{robvoigt,jurafsky}@stanford.edu

Abstract

We give an algorithm for disambiguating generic versus referential uses of second-person pronouns in restaurant reviews in Chinese. Reviews in this domain use the 'you' pronoun 你 either generically or to refer to shopkeepers, readers, or for self-reference in reported conversation. We first show that linguistic features of the local context (drawn from prior literature) help in disambigation. We then show that document-level features (n-grams and document-level embeddings)— not previously used in the referentiality literature— actually give the largest gain in performance, and suggest this is because pronouns in this domain exhibit 'one-sense-per-discourse'. Our work highlights an important case of discourse effects on pronoun use, and may suggest practical implications for audience extraction and other sentiment tasks in online reviews.

1 Introduction and Task Description

Detecting whether a given entity is referential is an important question in computational discourse processing. Linguistic features in the local context of a given mention have been successfully used for determining whether a second-person pronoun (you) in dialogue is referential (Gupta et al., 2007b; Frampton et al., 2009; Purver et al., 2009). The related task of anaphoricity detection is an important subtask of coreference resolution (Ng and Cardie, 2002; Ng, 2004; Luo, 2007; Zhou and Kong, 2009; Recasens et al., 2013).

In this paper we consider the task of audience identification in review texts, using restaurant reviews written in Chinese. Our task is to disambiguate a mention of the Chinese second-person pronoun 你 (*ni*, "you") into the following four labels that we found to occur commonly in reviews:

Generic
饮品只有雪碧和可乐，而且要点才拿给你
For drinks they only have Sprite and Coke, and you have to order before they'll give them to you.

Referential - Shop
这么好的服务下次还来你家哦
With such good service, I'll definitely come back to your shop next time!

Referential - Reader
不信你们去试你们会终身遗憾！
Go and try it if you don't believe me - your whole body will feel regret!

Referential - Writer / Self
店员说"你们就只要钵钵鸡？"
The shop employee said, "You only want the stone-bowl chicken?"

We aim to gain insight into the linguistics of narrative by distinguishing the types of discourse contexts in which different referential senses are found. Restaurant reviews provide an important new test case, and resolving who a reviewer wants to address could have important implications for coreference resolution or sentiment analysis of reviews, as well as downstream tasks like information extraction.

2 Related Work

A number of closely related earlier papers have focused on disambiguating 'you' in English. Gupta et al. (2007b) annotated the Switchboard corpus of telephone dialogue, showing that features based on specific lexical patterns, adjacent parts-of-speech, punctuation, and dialog acts are sufficient to achieve performance of 84.39% at the binary generic/referential prediction task. Gupta et al. (2007a) show that similar features generalize to addressee prediction for multi-party in-

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 314–319,
Beijing, China, July 26-31, 2015.

teractions significantly better than a simple baseline. Frampton et al. (2009) combine discourse features with acoustic and visual information for four-way interactions to resolve participant reference, and in the same setting Purver et al. (2009) employ cascaded classifiers that first establish referentiality and then attempt to resolve the referent. They show that utterance-level lexical features help, suggesting that different uses of 'you' are associated with distinct vocabularies.

Reiter and Frank (2010) investigate the more general question of identifying genericity for noun phrases, showing the usefuleness of linguistic features such as syntactic dependency relations. Similar local structural cues like phrase-structure positioning, head word identity, and distance to surrounding clauses have been used as features in machine learning approaches for anaphoricity detection as one stage in a coreference resolution (Kong and Zhou, 2010; Zhou and Kong, 2011; Kong and Ng, 2013).

Prior work has also shown improvements in performance in the dialogue domain from incorporating features having to do with acoustic prosody, gaze, and head movements (Jovanović et al., 2006; Takemae and Ozawa, 2006; Gupta et al., 2007b; Frampton et al., 2009). Of course in the review domain we have no access to such information; as we'll see, however, we can exploit other unique properties of reviews to make up for this lack.

3 Data

We scrape reviews from *dianping.com*, a Chinese-language restaurant review site, from the ten cities with the most reviews. We randomly sample 750 restaurants within each city and randomly sample reviews of those restaurants.

We scraped 346,381 reviews, including all associated metadata (city, restaurant category, and cost) for each restaurant, as well as the provided ratings (service, taste, ambience, and overall stars) for each review. Of these reviews only 6,704 (less than 2%) have the second-person pronominal character *ni*, highlighting another particular interest of this task: explicit second-person pronominals are quite rare in Chinese, at least in this genre, making the reviews in which they appear linguistically marked.

Summary statistics for this dataset are given in Table 1. We release all our data and annotations at `nlp.stanford.edu/robvoigt/nis`.

3.1 Preprocessing

We apply the Stanford CRF Word Segmenter (Tseng et al., 2005) to segment the text of each review into words, and use simple heuristics based on whitespace and punctuation to extract sentences or sentence fragments. The Stanford Parser (Klein and Manning, 2003; Levy and Manning, 2003) is then run on each extracted sentence or fragment containing a *ni* to produce a dependency graph and set of part-of-speech (POS) tags for later use in feature extraction.

3.2 Annotation

We hand-annotated 701 examples of *ni* tokens (including both singular and plural cases), placing them into one of seven categories: generic, writer-referential, reader-referential, shop-referential, idiomatic, non-"you", and other. The idiomatic and non-"you" cases are commonly comprised of set phrases such as 你好 (*nihao*, "hello") or 迷你 (*mini*, "mini") and are therefore relatively trivial to filter; and the "other" class is both rare and varied, including cases such as direct reference to prior review-writers.

We therefore only consider the generic and large-class referential cases, leaving us with 636 examples for our task; the distribution of annotated *ni*s is shown in Table 2.

The approximately half-and-half split between generic and referential tokens is surprisingly similar to that found by studies on English dialogue like Gupta et al. (2007b), in spite of the large divergence in language and genre.

We also found an unexpected word-sense property of second-person pronouns in this genre: of the 122 annotated reviews which contain more than one *ni*, 83.6% use *ni* with the same sense in each occurrence in the review, recalling the *one-sense-per-discourse* hypothesis of Gale et al. (1992). Finding that this discourse property—normally predicated of word-sense in common nouns—occurs in pronouns suggests the use of features of the entire discourse in this task.

4 Features

We consider two primary types of features: *"local"* and *"discourse"*.

4.1 Local Features

"Local" features model textual and linguistic properties of the immediate context of a given *ni*

SUBSET	small REVIEWS	CHARACTERS	WORDS	CHARS / REVIEW	WORDS / REVIEW
Total	346,381	15,010,375	10,112,722	43.33	29.20
Containing *ni*	6,704	1,099,597	748,683	164.02	111.68

Table 1: Summary statistics for the dataset collected for this paper; 701 cases of *ni* in 472 documents were annotated.

TYPE	ADDRESSEE	COUNT
Generic	-	296
Referential	Shop	256
Referential	Reader	48
Referential	Writer	36
Idiomatic	-	25
Non-"you"	-	26
Other	-	14

Table 2: Distribution of relevant types of 你 (*ni*, "you") in our annotated data.

mention, and were drawn from the large literature on referentiality, anaphoricity, and singleton-detection:

Word Identity This feature simply encodes the word-segmented identity of the word in which the current *ni* token is found, capturing cases such as the second-person plural 你们 (*nimen*, "you [plural]").

Adjacent POS Tags Following Gupta et al. (2007b), we include POS tag features for the single words immediately following and preceeding the *ni* token.

Dependencies We include binary features for the presence or absence of lexicalized dependency relations in which the given *ni* participates. As an example, for the phrase 你要推销菜 ("if you want to sell dishes"), we extract a feature for NSUBJ(推销, 你) – *you* is the subject of the verb *sell*.

Lexical Context This feature set fires binary features for the presence or absence of words in the vocabulary within a three-word window on either side of the given *ni* token.

4.2 Discourse Features

The *"discourse"* category considers features that characterize the entire review, capturing the intuition that the classic one-sense-per-discourse property is likely to hold for a given review, so we expect that features on the entire text of the review will be relevant for prediction.

This is a novel contribution of this work: we propose that in certain contexts (such as reviews), referentiality resolution can be interpreted in part as a text classification task.

Review N-grams These are binary features for the presence or absence of n-grams in the entire text of the review. We found that using a larger *n* than 1 caused overfitting on our relatively small dataset and reduced performance; therefore, results are reported using unigram features.

Review Vector Embedding To see if we can induce higher-level representations of the review text than simply binary n-gram features, we also train a document-level distributed vector representation (Le and Mikolov, 2014) on the entire corpus of reviews using the "doc2vec" implementation in GENSIM (Řehůřek and Sojka, 2010), and include 200 vector features per review: a 100-dimensional embedding learned on the entire document, as well as a 100-dimensional average embedding calculated by averaging the vectors for each word in the document. In experiments we found using both the document and the average vectors combined resulted in higher performance than either alone, so we report results in this setting.

Metadata In addition to discourse features, we also included features that encode the category, city, and estimated cost for each restaurant, as well as the service, taste, environment, and overall star rank ratings associated with a given review on a 5-point scale.

5 Experiments

We tested the effectiveness of these features at predicting genericity and reference for each *ni* token with multinomial logistic regression, as implemented in SCIKIT-LEARN (Pedregosa et al., 2011). We used two classification settings: a binary prediction of whether a given *ni* is referential or not, and a four-way prediction including distinctions between the three annotated referential targets. The results for each task are shown in Table 3.

In each case, we compare the performance of all local and discourse features, as well as several relevant subsets. One question we aim to address is whether our discourse-level n-gram and embed-

	Features	Binary	Four-way
Local	Baseline	53.44%	46.56%
	Word ID	67.19%	62.19%
	+ POS	74.84%	64.38%
	+ Deps	75.00%	69.38%
	+ Context	78.44%	72.19%
Discourse	N-grams	81.72%	77.03%
	Vectors	74.84%	66.56%
	N-grams + Vectors	81.25%	76.72%
Mixed	Local + N-grams	84.38%	79.84%
	Local + Vectors	86.56%	78.91%
	Local + Discourse	85.78%	80.63%
All	Local + Discourse + Meta	**88.21%**	**81.45%**

Table 3: Average ten-fold cross-validation classification accuracy for different feature sets on two tasks. "Local" refers to all feature sets described in Section 4.1. Binary distinguishes generic and referential *ni*, Four-way distinguishes between generic and three referential senses.

Feature	Binary		Four-way	
	estimate	*p*	*estimate*	*p*
ID	2.5%	***	1.9%	***
POS	1.0%	.	1.0%	.
Deps	0.2%		0.6%	
Context	4.2%	***	4.2%	***
N-grams	6.6%	***	7.8%	***
Vectors	3.8%	***	3.8%	***
Metadata	2.7%	***	2.6%	***

Table 4: Meta-analysis results for both tasks: effect size estimates from linear regressions ($n = 127$) predicting cross-validation scores from feature set. the *p* column denotes statistical significance; . is $p < 0.1$ and *** is $p < 0.001$.

ding features contribute similar information, so we test them both separately and together. We compare our results to a baseline of choosing the most common class for either task.

We train and test models with ten-fold cross-validation. In each fold, we use 80% of the data for training, 10% for development, and 10% for testing. For each feature set, we set the $l2$ regularization strength as a hyperparameter based on average cross-validation accuracy on the development data in each fold. All reported results are average cross-validation accuracy at that regularization strength on the test set in each fold.

5.1 Meta-analysis

To better understand the effectiveness of each feature set for this task, we perform a full ablation study by training a classifier on all 127 ($2^7 - 1$, ignoring the empty set) possible combinations of our 7 feature sets, and run a linear regression predicting the classification score from the feature sets used. This allows us to obtain estimates of the effect size and statistical significance for each set of features with reference to all the others. These results are shown in Table 4.

6 Discussion

These results show that on the task of detecting genericity and reference for second-person pronouns in our annotated set of Chinese-language restaurant reviews, both discourse-level features as well as local, contextual features significantly impact classification performance.

Simple word identity features alone already provide surprising performance: the classifier learns that the singular *ni* is more likely to be generic while the plural 你们 often refers to people affiliated with the shop.

While local features alone achieve respectable performance (78.44% for binary genericity detection and 72.19% for four-way classification), we show that in the review context significant gains can be made from using a combination of local and discourse-level features, exploiting discourse-level indicators of referentiality and the fact that a one-sense-per-discourse assumption tends to hold with regards to the use of *ni*.

Analysis of learned feature weights in our highest-performing model also provides some interesting social insights. Reviews with a high overall star rank were more likely to use generic *ni*, and reviewers who thought highly of the restaurant's service as indicated by their quality-of-service rating were more likely to use reader-directed referential *ni*.

Reviews with shop-directed referential *ni* were likely to use emotive sentence-final particles like 啊 (*a*), exclamation points, and question marks, just as question marks were among the strongest indicators of referential uses in the English "you"s in Gupta et al. (2007b). We also found that other pronouns like 我 (*wo*, "I") and 我们 (*women*, "we"), as well as words of temporal sequencing 第一 (*diyi*, "the first"), 又 (*you*, "again"), and 次 (*ci*, "[one] time") receive high weights for referential classes.

Combined with the observation that reviews containing *ni* simply tend to be much longer than those without (see Table 1), these results suggest a link to the narrative work of Jurafsky et al. (2014),

who characterize negative reviews as narrative expositions of an individual bad experience.

For example, consider the following review containing a referential *ni*:

> 菜品，份量都不错。环境更没得说。但我们中午去的晚，没想到人家先关灯，后又关空调。想问下你们省电了，是想证明我们吃饭可以不用给钱吗?
>
> The food and quantity was fine. The ambience need not be mentioned. But in spite of having been a bit late for lunch, we wouldn't have imagined you'd first turn off the lights, and then turn off the air conditioner. I'd like to ask: saving money on electricity like this, do you mean to imply that there's no need for us to pay for our meal?

While the immediate context suggests a referential interpretation (想问下你们省电了, literally "want to ask you [plural], saving electricity"), it is only when this mention is connected to elements of the entire discourse (the sequence of events, the first-person pronouns) that it becomes completely clear first that the mention is referential and second that it refers to the shop owner.

Furthermore, we found that when combined with local features, features derived from distributed representations of each document perform at least as well for this task as document-level n-grams, but at a much lower dimensionality. This suggests that these embeddings do successfully encode the information necessary to reproduce document-level distinctions in discourse types, such as between the personal narratives that often surround referential uses of *ni* and the abstract descriptions of generic uses.

Our meta-analysis shows that more linguistically motivated local features such as POS tags and dependency relations are substantially overshadowed in effectiveness by lexical and discourse features, although this may be due in part to reduced performance of these automatic taggers on the more colloquial language in online reviews.

Finally, this work challenges prior claims that spoken language is "more complex" than other genres with regards to referentiality. On the contrary: whereas in a spoken discourse the potential addressees are by default the participants, web texts such as the reviews studied here have no such default, and may include complex, creative, and domain-specific deictic reference that can be important for computational systems to address.

References

Matthew Frampton, Raquel Fernández, Patrick Ehlen, Mario Christoudias, Trevor Darrell, and Stanley Peters. 2009. Who is you?: combining linguistic and gaze features to resolve second-person references in dialogue. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 273–281. Association for Computational Linguistics.

William A Gale, Kenneth W Church, and David Yarowsky. 1992. One sense per discourse. In *Proceedings of the workshop on Speech and Natural Language*, pages 233–237. Association for Computational Linguistics.

Surabhi Gupta, John Niekrasz, Matthew Purver, and Daniel Jurafsky. 2007a. Resolving "you" in multiparty dialog. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 227–30.

Surabhi Gupta, Matthew Purver, and Dan Jurafsky. 2007b. Disambiguating between generic and referential you in dialog. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 105–108. Association for Computational Linguistics.

Natasa Jovanović, Anton Nijholt, et al. 2006. Addressee identification in face-to-face meetings. Association for Computational Linguistics.

Dan Jurafsky, Victor Chahuneau, Bryan R Routledge, and Noah A Smith. 2014. Narrative framing of consumer sentiment in online restaurant reviews. *First Monday*, 19(4).

Dan Klein and Christopher D Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

Fang Kong and Hwee Tou Ng. 2013. Exploiting zero pronouns to improve chinese coreference resolution. In *EMNLP*, pages 278–288.

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. Association for Computational Linguistics.

Quoc V Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*.

Roger Levy and Christopher Manning. 2003. Is it harder to parse chinese, or the chinese treebank? In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 439–446. Association for Computational Linguistics.

Xiaoqiang Luo. 2007. Coreference or not: A twin model for coreference resolution. In *HLT-NAACL*, pages 73–80.

Vincent Ng and Claire Cardie. 2002. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

Vincent Ng. 2004. Learning noun phrase anaphoricity to improve coreference resolution: Issues in representation and optimization. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 151. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Matthew Purver, Raquel Fernández, Matthew Frampton, and Stanley Peters. 2009. Cascaded lexicalised classifiers for second-person reference resolution. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 306–309. Association for Computational Linguistics.

Marta Recasens, Marie-Catherine de Marneffe, and Christopher Potts. 2013. The life and death of discourse entities: Identifying singleton mentions. In *HLT-NAACL*, pages 627–633.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. `http://is.muni.cz/publication/884893/en`.

Nils Reiter and Anette Frank. 2010. Identifying generic noun phrases. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 40–49. Association for Computational Linguistics.

Yoshinao Takemae and Shinji Ozawa. 2006. Automatic addressee identification based on participants' head orientation and utterances for multiparty conversations. In *Multimedia and Expo, 2006 IEEE International Conference on*, pages 1285–1288. IEEE.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter. In *Proceedings of the fourth SIGHAN workshop on Chinese language Processing*, volume 171.

Guodong Zhou and Fang Kong. 2009. Global learning of noun phrase anaphoricity in coreference resolution via label propagation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 978–986. Association for Computational Linguistics.

Guodong Zhou and Fang Kong. 2011. Learning noun phrase anaphoricity in coreference resolution via label propagation. *Journal of Computer Science and Technology*, 26(1):34–44.