Using Discourse Features for Referring Expression Generation

Emily Jamison
Department of Linguistics
The Ohio State University
Columbus, OH 43210, USA
jamison@ling.osu.edu

Abstract

Referring expression generation (REG) is the task of choosing a noun phrase that refers to an extra-linguistic entity and inserting the noun phrase into a text. REG is helpful for tasks such as text generation and information summarization (Kan et al., 2001). We present a system for referring expression generation that mines a number of discourse-based feature functions from text and uses a maximum entropy classifier (Le, 2004) to choose a referring expression from a list of possibilities. Our results show varying effectiveness of the different discourse-based predicates. We also demonstrate two different uses of the maximum entropy classifier: a binary classification method and an n-class classification method.

1 Introduction

Referring expression generation (*REG*) is the task of inserting noun phrases that refer to a mentioned extra-linguistic entity into a text. A referring expression is a phrase that refers to an extra-linguistic entity. A single entity can be referred to using a number of phrases in a discourse. For example, in figure (1), *Felix* could also be referred to as *the cat*, *he*, or *the neighbor's pet*.

(1)
Felix chased a chipmunk across the yard.

Some expressions that refer to an entity are limited to certain discourse contexts. For example, in figure (2a), *Felix* is first refered to by name and subsequently with a pronoun and an empty expression. In figure (2b), *Felix* is always referred to by name, which sounds less natural, since later mentions of an entity within a discourse are often reduced to a pronominal form, or eliminated altogether.

(2a)
Felix ran through the violets. *He* dodged under the lilac and _ even followed the chipmunk right up the cherry tree.

(2b)
Felix ran through the violets. *Felix* dodged under the lilac and *Felix* even followed the chipmunk right up the cherry tree.

The goal of REG is to automatically determine which referring expression is the most appropriate phrase to refer to an entity, given the surrounding text. REG is helpful for tasks such as text generation and information summarization (Kan et al., 2001).

We present a system for referring expression generation that mines a number of discourse-based feature functions from the text and uses a maximum entropy classifier (Le, 2004) to choose a referring expression from a list of possibilities. We demonstrate our system on the GREC corpus (Belz and Gatt, 2008), a collection of introductory texts from Wikipedia in which the main topic of each text (the main subject referent, or *MSR*) has been replaced with a list of possible referring expressions (REs). The system's performance is evaluated by determining whether the predicted RE is of the same type (proper name, common noun, pronoun, or empty) as the original RE.

2 Corpus

The GREC corpus (Belz and Gatt, 2008) is a collection of introductory texts from Wikipedia in which each mention of the main topic of each text (*MSR*) has been replaced with a list of possible referring expressions (REs). The list of possible referring expressions includes all the expressions in the text used to refer to the topic, as well as several automatically-generated expressions appropriate for the topic. Figure (3) shows a sample text from the corpus; each mention of the MSR, *Albania*, has been replaced with a list of possible REs. The correct RE from the original article is bolded.

> (3) Albania
> {*Albania, Albania itself, _, it, it itself, that, that itself,* **the Republic of Albania**, *the Republic of Albania itself, the country, the country itself, which, which itself*}
> (Albanian: Republika e Shqipris, IPA [IPA: [publika cipis]]) is a Balkan country in Southeastern Europe.
> {*Albania, Albania itself, _,* **it**, *it itself, that, that itself, the Republic of Albania, the Republic of Albania itself, the country, the country itself, which, which itself*}
> borders Montenegro to the north, the Serbian province of Kosovo to the northeast, the Republic of Macedonia in the east, and Greece in the south.
> {*Albania, Albania itself, _,* **it**, *it itself, that, that itself, the Republic of Albania, the Republic of Albania itself, the country, the country itself, which, which itself*}
> has a coast on the Adriatic Sea to the west and a coast on the Ionian Sea to the southwest. Despite having a troubled history,
> {*Albania, Albania itself, _, it, it itself, that, that itself, the Republic of Albania, the Republic of Albania itself,* **the country**, *the country itself, which, which itself*}
> has been classified as an emerging democracy since the 1990s.

In the GREC corpus, each RE is annotated with the following feature functions: expression type (name, common noun, pronoun, empty), whether or not the feature is emphatic, the type of the RE's head (nominal, pronoun, rel-pron), and case of the head (nominative, accusative, genitive, plain). In addition, each mention of the MSR is annotated with the following feature functions: semantic category (city, country, river, mountain, person) and syntactic category (np-obj, np-subj, subj-det). Texts in the GREC corpus are from five categories: cities, countries, mountains, rivers, and people. Annotated REs include subject, objects, and genetive subject-determiners. These categories are known to have high inter-annotator agreement in identifying mentions (Belz and Gatt, 2008).

3 Predicates

We created five groups of predicates, in addition to the predicates available with the corpus. All predicates can be used with the binary classification method; only non-RE-specific predicates can be used with the n-class classification method.

The following groups of predicates were used:

- CDE: detection of competing discourse entities
- TITLE: comparison of the RE and the article's title
- ORDER: mention order
- DIST: distance between the current mention and the previous mention
- PREV: whether the RE was used previously in the text

We assumed that the presence of a competing discourse entity would reduce the likelihood of pronouns or empty forms being used in that case (CDE). To detect competing discourse entities, separate predicates looked for the definite determiner 'the', as well as gender matching pronouns, in the whole text, the text preceding the mention, or the text between the current mention and the most recent mention.

Predicates involving the article title (TITLE) were included on the basis that the most common RE form would be used in the title. The three TITLE predicates compared character length for the title and the RE, and checked whether the RE was the same as or contains the title.

The initial RE and subsequent mentions are likely to have a differing type distribution. Specifically, we expect that an initial mention might be more likely to use a proper name, while later mentions may be pronouns or common nouns. We included a mention order predicate that identifies the mention as first, second, middle, or last in the text (ORDER).

If the text distance to the previous mention is large, the mention is less likely to be a pronoun. We included two discretized distance predicates (DIST) to reflect this: one predicate measured distance in words (with bins of 0-5, 6-12, and 13+ words) and the other predicate measured distance in sentence boundaries (with values of 0, 1, or 2+ boundaries).

Since a previous occurrence of an RE may influence which REs are used later in the text as well, the predicate PREV noted whether or not the RE under consideration had been previously used in the text.

4 Maximum Entropy Classifier

We used Le's Maximum Entropy Modeling Toolkit (Le, 2004). We generated feature functions from our predicates, combining a property of the syntactico-semantic and discourse context of an RE with a label. Our feature functions f_i were used to train a maximum entropy classifier (Berger et al., 1996) (Le, 2004) that assigns a probability to an RE re, or mention, given context c_x as follows:

$$p(re|\, c_x) = Z(c_x) \exp \sum_{i=1}^{n} \lambda_i f_i(c_x, re)$$

where $Z(c_x)$ is a normalizing sum and the λ_i are the parameters (feature weights) learned.

Two classification systems were used: binary and n-class. With the binary method, the classifier estimates the likelihood of a possible referring expression's correct insertion into the text, and inserts the RE with the highest 'yes' probability from the list of possible REs for a given mention. With the n-class method, the mention itself is classified according to type of referring expression (proper name, common noun, pronoun, empty) and a RE of the proper type is chosen. Figure (4) illustrates the binary method, while figure (5) illustrates the n-class method. The system-chosen RE is bolded.

(4): Binary method
{**Felix** yes=.58 no=.42; *he* yes=.36 no=.64; *the cat* yes=.55 no=.45; *the neighbor's pet* yes=.49 no=.51} chased a chipmunk across the yard.

(5): N-class method
{name=.33 commonnoun=.21 pronoun=.40 empty=.06 Selected:**he**} chased a chipmunk across the yard.

Predicates Used	Type Accuracy
most frequent ('pronoun')	40.40%
GREC predicates	50.91%
GREC + CDE	51.07%
GREC + TITLE	50.30%
GREC + ORDER	61.13%
GREC + DIST	51.83%
GREC + PREV	51.07%
all except PREV	58.54%
all incl. PREV	57.62%

Table 1: Results with binary classification.

Predicates Used	Type Accuracy
most frequent ('pronoun')	40.40%
GREC predicates	51.83%
GREC + CDE	54.27%
GREC + TITLE	N.A.
GREC + ORDER	62.65%
GREC + DIST	55.95%
GREC + PREV	N.A.
all except PREV	62.50%
all incl. PREV	N.A.

Table 2: Results with n-class classification.

Maximum entropy learners cannot learn conjunctions of multiple features, so a feature combinator was implemented that created pairs of features associated with a binary value. The feature combinator was used for both maximum entropy methods.

5 Results

Our results are shown in tables 1 and 2. PREV has been separated from one of the all-inclusive predicate sets because it is unique in that it assumes perfect knowledge of the REs that had been used in previous discourse, which is an assumption unlikely to hold in a real-life, continuous generation context.

The most helpful of the groups of predicates added to the GREC information was ORDER (Binary: 50.91% to 61.13% over standard GREC information; N-class: 51.83% to 62.65%). The least helpful group of predicates was TITLE (binary: 50.30% with these predicates and 50.91% without).

6 Discussion

A number of discourse-level predicates increased the type accuracy rate of referring expression generation in this experiment. A combination of our features increased the binary classification from 50.91% to 57.62%, and n-class classification from 51.83% to 62.50%.

The presence of competing discourse entities was expected to be highly influential in determining the form of the RE to be used. Specifically, if another discourse entity was present in the text immediately preceding the mention, a pronoun would not be used, to avoid referent ambiguity. This ambiguity is illustrated in figure (6a) with the use of a pronoun to refer to the cat when the mention was preceded by a competing discourse entitiy (the dog). Figure (6b) shows the same sentence with a non-ambiguous RE.

> (6a)
> The chipmunk had been spotted by both the cat and the dog. **It** gave chase to the rodent, nearly catching the poor animal at the fence.
>
> (6b)
> The chipmunk had been spotted by both the cat and the dog. **The cat** gave chase to the rodent, nearly catching the poor animal at the fence.

However, CDE only improved accuracy of the n-class system from 51.83% to 54.27%, and altogether failed to noticeably improve binary system accuracy. Subsequent investigation into the performance of CDE revealed that its only mildly improved accuracy may be rooted in inherent dilemmas of coreference annotation, as illustrated by the following example from the GREC corpus.

> (7): Alfred Nobel
> **Alfred Nobel** was a Swedish chemist, engineer, innovator, armaments manufacturer and the inventor of dynamite. [...] In **his** last will, **Alfred Nobel** used **his** enormous fortune to institute the Nobel Prizes.

Figure (7) shows a sample text from the GREC corpus. In addition to the marked REs, other unmarked phrases ("the inventor of dynamite") may also be refer to the same entity; a predicate identifying these unmarked REs as "competing discourse entities" may not provide useful information.

7 Future Work

In this study, we used our set of discourse-level predicates and a separate binary and an n-class maximum entropy classifier to select referring expressions. In future work, we plan to combine these two classification methods, first selecting all REs of the appropriate type and then ranking each previously-selected RE. Also, it is possible that the use of a corpus with more coreference annotation would result in improved performance of CDE. For example, the ACE corpus[1] has a greater variety of annotated entities and might provide additional data for such investigations.

Acknowledgments

The author is indebted to Dennis Mehay for his explanation of the maximum entropy classifier. The author would also like to express appreciation to Yannick Versley and Steve Boxwell, as well as 3 anonymous reviewers, for their insightful comments on earlier versions of this paper.

References

Anya Belz and Albert Gatt. 2008. *REG Challenge 2008: Participants Pack*. http://www.nltg.brighton.ac.uk/research/reg08/.

A. L. Berger, S. D. Pietra, and V. D. Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Min-Yen Kan, Kathleen R. McKeown, and Judith L. Klavans. 2001. Applying natural language generation to indicative summarization. *EWNLG '01: Proceedings of the 8th European workshop on Natural Language Generation*.

Zhang Le. 2004. *Maximum Entropy Modeling Toolkit for Python and C++*. http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html.

[1]Mitchell et al. (2003) *ACE-2 Version 1.0* LDC2003T11