Learning Contrastive Connectives in Sentence Realization Ranking

Crystal Nakatsu
Department of Linguistics
The Ohio State Univeristy
Columbus, OH, USA
cnakatsu@ling.osu.edu

Abstract

We look at the average frequency of contrastive connectives in the SPaRKy Restaurant Corpus with respect to realization ratings by human judges. We implement a discriminative n-gram ranker to model these ratings and analyze the resulting n-gram weights to determine if our ranker learns this distribution. Surprisingly, our ranker learns to avoid contrastive connectives. We look at possible explanations for this distribution, and recommend improvements to both the generator and ranker of the sentence plans/realizations.

1 Introduction

Contrastive discourse connectives are words or phrases such as *however* and *on the other hand*. They indicate a contrastive discourse relation between two units of discourse. While corpus-based studies on discourse connectives usually look at naturally occurring human-authored examples, in this study, we investigate the set of connectives used in the automatically generated SPaRKy Restaurant Corpus[1]. Specifically, we consider the relationship between connective usage and judges ratings, and investigate whether our n-gram ranker learns the preferred connective usage. Based on these findings and previous work on contrastive connectives, we present suggestions for modifying both the generator and the ranker in order to improve the generation of realizations containing contrastive connectives.

[1]We thank Marilyn Walker and her research team for making all of the MATCH system data available for our study, especially including the SPaRKy Restaurant Corpus.

2 Corpus Study

2.1 SPaRKy Restaurant Corpus

The SPaRKy Restaurant Corpus was generated by the MATCH Spoken Language Generator (Walker et al., 2007) which consists of a dialog manager, SPUR text planner (Walker et al., 2004), SPaRKy sentence planner (Walker et al., 2007), and RealPro surface realizer (Lavoie and Rambow, 1997).

The corpus contains realizations for 3 dialogue strategies:

- RECOMMEND (REC): recommend an entity from a set of entities
- COMPARE-2 (C2): compare 2 entities
- COMPARE-3 (C3): compare 3 or more entities

Each strategy contains 30 content plans from which either 16 or 20 sentence plans were generated by the SPaRKy sentence plan generator. 4 sentence plans were discarded due to duplication upon realization, totaling 1756 realizations in the corpus.[2]

A content plan consists of several assertions and the relations which hold between them. Content plans from the RECOMMEND strategy exclusively employ the Rhetorical Structure Theory (RST) (Mann and Thompson, 1987) relation JUSTIFY while those from COMPARE-2 use CONTRAST and ELABORATION. COMPARE-3 content plans consists mostly of CONTRAST and ELABORATION relations, though some use only JUSTIFY. In addi-

[2]The total number of realizations reported here is inconsistent with the information reported in (Walker et al., 2007). In corresponding with the authors of that paper, it is unclear why this is the case; however, the difference in reported amounts is quite small, and so should not affect the outcome of this study.

Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue, pages 76–79,
Columbus, June 2008.

Strategy	Alt #	Rating	Rank	Realization
C2	3	3	7	Sonia Rose has very good decor but Bienvenue has decent decor.
	7	1	16	Sonia Rose has very good decor. On the other hand, Bienvenue has decent decor.
	8	4.5	13	Bienvenue has decent decor. Sonia Rose, on the other hand, has very good decor.
	10	4.5	5	Bienvenue has decent decor but Sonia Rose has very good decor.
	11	1	12	Sonia Rose has very good decor. However, Bienvenue has decent decor.
	13	5	14	Bienvenue has decent decor. However, Sonia Rose has very good decor.
	14	5	3	Sonia Rose has very good decor while Bienvenue has decent decor.
	15	4	4	Bienvenue has decent decor while Sonia Rose has very good decor.
	17	1	15	Bienvenue's price is 35 dollars. Sonia Rose's price, however, is 51 dollars. Bienvenue has decent decor. However, Sonia Rose has very good decor.

Figure 1: Some alternative [Alt] realizations of SPaRKy sentence plans from a COMPARE-2 [C2] plan, with averaged human ratings [Rating] (5 = highest rating) and ranks assigned by the n-gram ranker [Rank] (1 = top ranked).

tion, the SPaRKy sentence plan generator adds the INFER relation to assertions whose relations were not specified by the content planner.

During the sentence planning phase, SPaRKy orders the clauses and combines them using randomly selected clause-combining operations. During this process, a clause-combining operation may insert 1 of 7 connectives according to the RST relation that holds between two discourse units (i.e. inserting *since* or *because* for a JUSTIFY relation; *and*, *however*, *on the other hand*, *while*, or *but* for a CONTRAST relation; or *and* for an INFER relation).

After each sentence plan is generated, it is realized by the RealPro surface realizer and the resulting realization is rated by two judges on a scale of 1-5, where 5 is highly preferred. These ratings are then averaged, producing a range of 9 possible ratings from $\{1, 1.5, ..., 5\}$.

2.2 Ratings/Connectives Correlation

From the ratings of the examples in Figure 1, we can see that some of the SPaRKy sentence plan realizations seem more natural than others. Upon further analysis, we noticed that utterances containing many contrastive connectives seemed less preferred than those with fewer or no contrastive connectives.

To quantify this observation, we calculated the average number of connectives (ave_{c_i}) used per realization with rating i, using $ave_{c_i} = Total_{c_i}/N_{r_i}$, where $Total_{c_i}$ is the total number of connectives in realizations with rating i, and N_{r_i} is the number of realizations with rating i.

We use Pearson's r to calculate each correlation (in each case, $df = 7$). For both COMPARE strategies (represented in Figure 2(a) and 2(b)), we find a significant negative correlation for the average number of connectives used in realizations with a given rating (C2: $r = -0.97, p < 0.01$; and C3: $r = -0.93$, $p < 0.01$). These correlations indicate that judges' ratings decreased as the average frequency of the connectives increased.

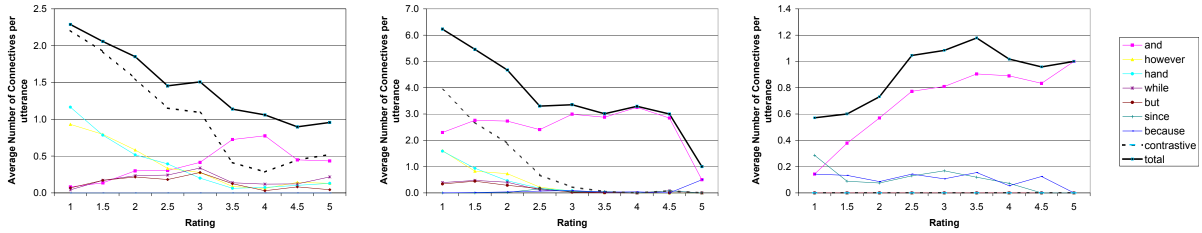
Further analysis of the individual correlations used in the comparative strategies show that there is a significant negative correlation for *however* (C2: $r = -0.91$, $p < 0.01$; and C3: $r = -0.86$, $p < 0.01$) and *on the other hand* (C2: $r = -0.89$, $p < 0.01$; and C3: $r = -0.84$, $p < 0.01$) in both COMPARE strategies. In addition, in COMPARE-3, the frequencies of *while* and *but* are also significantly and strongly negatively correlated with the judges' ratings ($r = -0.86$, $p < 0.01$ and $r = -0.90$, $p < 0.01$, respectively), though there is no such correlation between the use of these connectives and their ratings in COMPARE-2.

Added together, all the contrastive connectives show strong, significant negative correlations between their average frequencies and judges' ratings for both comparative strategies (C2: $r = -0.93$, $p < 0.01$; C3:$r = -0.88$, $p < 0.01$).

Interestingly, unlike in the COMPARE strategies, there is a positive correlation ($r = 0.73, p > 0.05$) between the judges' ratings and the average frequency of all connectives used in the RECOMMEND strategy (see Figure 2(c)). Since this strategy only uses *and*, *since*, and *because* and does not utilize any contrastive connectives, this gives further evidence that only contrastive connectives are dispreferred.

2.3 N-gram Ranker and Features

To acertain whether these contrastive connectives are being learned by the ranker, we re-implemented the n-gram ranker using SVM-light (Joachims,



(a) Compare 2 (b) Compare 3 (c) Recommend

Figure 2: Correlation Graphs: The thick solid line indicate the correlation of all the connectives summed together, while the thick dashed line indicates the correlation of the 4 contrastive connectives summed together.

Strategy	however	o.t.o.h	while	but	all contrastives
C2	25.0%	25.0%	0.9%	2.7%	53.6%
C3	9.9%	10.9%	0.0%	3.1%	24.0%

Table 1: The proportion of the 20% most negatively weighted features for all contrastive connectives.

2002). As in Walker et. al (2007), we first prepared the SPaRKy Restaurant Corpus by replacing named entity tokens (e.g numbers, restaurant names, etc.) with their corresponding type (e.g. NUM for *61*), and added BEGIN and END tokens to mark the boundaries of each realization. We then trained our ranker to learn which unigrams, bigrams, and trigrams are associated with the ratings given to the realizations in the training set.

Although we implemented our ranker in order to carry out an error analysis on the individual features (i.e. n-grams) used by the ranker, we also found that our n-gram ranker performed comparably (REC: 3.5; C2: 4.1; C3: 3.8)[3] to the full-featured SPaRKy ranker (REC: 3.6; C2: 4.0; C3: 3.6) out of a possible best (human-performance) score of (REC: 4.2; C2: 4.5; C3: 4.2).

Using a perl script[4], we extracted feature weights learned by the ranker from the models built during the training phase. After averaging the feature weights across 10 training partitions, we examined the top 20% (C2:112/563 features; C3: 192/960 features) most negatively weighted features in each strategy to see whether our ranker was learning to avoid contrastive connectives. Table 1 shows that features containing contrastive connectives make up 53.6% of the 20% most negatively weighted features in COMPARE-2 and 24.0% of the 20% of the most negatively weighted features used in COMPARE-3. Interestingly, COMPARE-2 features that contained either *however* or *on the other hand* (*o.t.o.h*) make up the bulk of the contrastive connectives found in the negatively weighted features, mirroring the results of the correlations for COMPARE-2. This indicates that the discriminative n-gram ranker learns to avoid using contrastive connectives.

[3]These scores were calculated using using the TopRank evaluation metric (Walker et al., 2007).

[4]written by Thorsten Joachims

3 Contrastive Connectives Usage

3.1 Usage Restrictions

Previous work on contrastive connectives have found that these connectives often have different restrictions on their location in the discourse structure, with respect to maintaining discourse coherance (Quirk et al., 1972; Grote et al., 1995).

Quirk et. al. (1972) classifies *however* and *on the other hand* as subordinating conjuncts, a class of connectives that do not allow their clauses to be reordered without changing the perlocutionary force of the sentence (e.g. contrast C2: Alts # 11 & 13 in Figure 1). In addition, *on the other hand* prompts readers to regard the 2nd clause as more important (Grote et al., 1995). Given that both *however* and *on the other hand* contain the same restrictions on clause ordering, it seems reasonable that they would pattern the same with respect to assigning clausal prominence. This predicts that if the human judges rated the SPaRKy realizations based on the expectation of a particular perlocutionary act (e.g., that the comparison highlights the restaurant with the best decor), they would prefer realizations where *however* or *on the other hand* were attached to the more desirable of the contrasted qualities. When we examine the SPaRKy realizations and ratings, this indeed seems to be the case – when the better property is ordered last, the realization was rated very highly (e.g. Alt 8 & 13 in Figure 1), but when the lesser property was ordered last, the realization was rated poorly (e.g. Alt 7 & 11 in Figure 1).

In contrast, *while* and *but* are not subordinating conjuncts and so are not subject to the clause ordering restriction. Thus, realizations with their contrasted clauses in either order should be rated similarly, and indeed, this is what we find in the corpus (e.g. Alts 3&10, and 14&15 in Figure 1).

3.2 Other Factors

In addition to clause order, another factor that may contribute to the awkwardness of *however* and *on the other hand* in some usages is that both of these connectives seem to be rather "grand" for these simple contrasts. Intuitively, these connectives seem to indicate a larger contrast than *while* and *but*, so when they are used to indicate small contrasts (e.g. contrasting only one quality), or contrasts close together on the scale (e.g. good vs. decent) instead of diametric opposites, they sound awkward. In addition, *however* and *on the other hand* may also be seeking "heavy" arguments that contain more syllables, words, or complex syntax. Lastly, human-authored comparisons, such as in this example from CNET.com:

> ...[it] has two convenient USB ports at the bottom of the front panel. Its beige predecessor, **on the other hand**, supplied these **only** on the back of the box.

seem to indicate that when our expectations of argument order are violated, the 2nd clause is often qualified by words such as "just" or "only", as if to acknowledge the flaunted preference.

4 Discussion and Future Work

Due to the poverty of highly rated instances of contrastive connective usage (particularly for *however* and *on the other hand*), our ranker learns to avoid these connectives in most situations. However, the ratings suggest that people do not dislike these contrastives unilaterally, but rather prefer them in specific usage patterns only. One way to combat this problem is to modify the sentence planner to take into account these semantic preferences for argument ordering when selecting a contrastive connective. This should produce a wider variety of candidates that observe this ordering preference, and thus provide the ranker with more highly rated candiates that use contrastive connectives. This is not to say that only candidates observing this preference should be generated, but merely that a wider variety of candiates should be generated so that the ranker has more opportunities to learn the restrictions surrounding the use of contrastive connectives.

As for the ranker, we can also identify features that are sensitive to these linguistic properties. Currently, n-gram features don't capture the semantic nuances such as argument order or the scalar distance between property values, so identifying features that capture this type of information should improve the ranker. Together, these improvements to both the quality of the generated candidate space and the ranking model should improve the accuracy of the top-rated/selected candidate.

References

B. Grote, N. Lenke, and M. Stede. 1995. Ma(r)king concessions in english and german. In *Proceedings of the Fifth European Workshop on Natural Language Generation.*, May.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proc. KDD*.

B. Lavoie and O. Rambow. 1997. A fast and portable realizer for text generation systems. In *Proceedings of the 5th. Conference on Applied Natural Language Processing*, pages 265–268, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

William C. Mann and Sandra A. Thompson. 1987. Rhetorical structure theory: A theory of text organization. Technical Report ISI/RS-87-190, University of Southern California Information Sciences Instuture.

R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik. 1972. *A Comprehensive Grammar of the English Language*. Longman.

M. A. Walker, S. J. Whittaker, A. Stent, P. Maloor, J. D. Moore, M. Johnston, and G Vasireddy. 2004. Generation and evaluation of user tailored responses in multimodal dialogue. *Cognitive Science*, 28(5):811–840.

M. Walker, A. Stent, F. Mairesse, and Rashmi Prasad. 2007. Individual and domain adaptation in sentence planning for dialogue. *Journal of Artificial Intelligence Research (JAIR)*, 30:413–456.