Contextual Influences on Near-Synonym Choice

Ehud Reiter and Somayajulu Sripada

{ereiter,ssripada}@csd.abdn.ac.uk
Dept of Computing Science, University of Aberdeen, UK

Abstract. One of the least-understood aspects of lexical choice in Natural Language Generation is choosing between near-synonyms. Previous studies of this issue, such as Edmonds and Hirst [4], have focused on semantic differences between near-synonyms, as analysed by lexicographers. Our empirical analysis of near-synonym choice in weather forecasts, however, suggests that other factors are probably more important than semantic differences. These include preferences and idiosyncrasies of individual authors; collocation; variation of lexical usage; and position of a lexeme in a text. Our analysis also suggests that when semantic differences do influence near-synonym choice, they may do so in an author-dependent manner. Thus, at least in our domain, 'context' (including author) seems to be more important than semantics when choosing between near-synonyms.

1 Introduction

Natural Language Generation (NLG) systems must choose the words used in a generated text. This task is called *lexical choice*. Lexical choice can be an especially difficult task when the NLG system wishes to communicate a meaning that can be expressed by several synonyms or near-synonyms; which of these should be used? For example, if the system wishes to communicate the time 1200, should it express this as *midday*, *noon*, *1200*, *late morning* or *early afternoon*?

In this paper we describe our work on near-synonym choice, which (unlike most previous work in this area) is based on corpus analysis and other empirical work. This work was motivated by a desire to develop more sophisticated representations of lexical meanings which could capture fine semantic nuances. In fact, though, our strongest finding is that the choice between near-synonyms is mostly determined by non-semantic factors, including the preferences and 'idiolect' of individual authors (which in some cases change over time); collocation; which words were previously used for similar concepts in the text; and the position of a lexeme in a text and sentence. At least in our domain, these 'contextual' factors seem to have more impact than subtle semantic differences on near-synonym choice.

2 Background

2.1 Lexical Choice

Lexical choice is the task of choosing words; it obviously is a key task in NLG. In this paper we focus on the choice of content (open-class) words. We also assume the three-stage pipeline architecture of Reiter and Dale [10], where lexical choice is part of the middle stage, *microplanning*. This means that the lexical choice module is given as input some specification of the information the NLG system wishes to communicate (this comes from content determination in the first pipeline stage); and produces as output a word or set of words in their root forms (morphology is done during the third stage, realisation). Finally, we assume that lexical choice is carried out on small chunks of information; that is, the lexical choice module is given information in small chunks, and is expected to choose a small number of words (often just one) for each chunk.

In summary, the task of lexical choice as we define it is to select a word or a small set of words that communicate a small piece of content information; for example, choosing *midday* to express the meaning 1200, as described above. For other perspectives on lexical choice, see Fawcett *et al* [5], who integrates lexical choice with realisation; Reiter [9] and Stone *et al* [17], who treat lexical choice as a pragmatic process of satisfying communicative goals; and Robin and McKeown [13], who treat lexicalisation as an optimisation process which chooses many words simultaneously.

Lexical choice from this perspective has two aspects, search and choice. Search is the task of finding those lexemes (or sets of lexemes) that communicate the desired information; choice is the task of choosing which of these candidates to actually use. Our focus here is on choice, and in particular the choice between near-synonyms.

Perhaps the best known previous work on choosing between near-synonyms is Edmonds and Hirst [4]. They assume that words are grouped into 'clusters' of near-synonyms. Each cluster has a core meaning or denotation, which all words in the cluster communicate. Near-synonyms in a cluster are distinguished by further 'peripheral' distinctions, which indicate fine shades of meaning or connotation which are compatible with the core meaning. These distinctions are based on a standard synonym dictionary [6]. Actual lexical choice is based on a matching and optimisation process, which searches for the near-synonym which most closely matches the current situation.

Edmonds and Hirst's work is impressive, but it is not empirically based in the sense that they did not themselves perform corpus analysis or psycholinguistic experiments (although presumably the definitions in the synonym dictionary they used had some empirical basis). They also did not look at the influence of the kind of contextual effects we discuss in this paper, probably because this information was not present in their synonym dictionary.

In an earlier paper [11], we discussed the fact that different people may associate different meanings with words. For example, our data showed that some people interpret *by evening* to mean 1800 (6PM), while others interpret

FORECAST 00-24 GMT, 18-Sep 2000 MONDAY
WIND(10M): SSE 10-14 RISING 24-28 IN THE MORNING THEN VEERING SSW EARLY AFTERNOON AND EASING 16-20 LATER
(50M): SSE 12-18 RISING 30-35 IN THE MORNING THEN VEERING SSW EARLY AFTERNOON AND EASING 20-25 LATER
SIG WAVE: 1.0-1.5 RISING 2.0-2.5, LOCALLY 2.5-3.0 FOR A TIME
MAX WAVE: 1.5-2.5 RISING 3.0-4.0, LOCALLY 4.0-5.0 FOR A TIME
WEATHER: RAIN CLEARING BY EVENING. CONTINUING RISK OF MIST AND FOG

Fig. 1. Extract from 5-day human-authored forecast issued on 16-Sep-00

Table 1. Wind (at 10m) extract from 16-Sep-00 NWP data file

day	hour	wind dir	wind speed
18-09-00	0	SSE	10
18-09-00	3	S	13
18-09-00	6	S	18
18-09-00	9	S	22
18-09-00	12	S	26
18-09-00	15	SSW	25
18-09-00	18	SW	23
18-09-00	21	SW	19
19-09-00	0	SSW	17

it to mean 0000 (midnight). This paper in part continues our exploration of individual differences in language usage, but it focuses on near-synonym choice instead of semantic interpretation.

2.2 SumTime-Mousam

Our empirical work was carried out using data from SumTime-Mousam [14]. SumTime-Mousam is an NLG system that produces textual weather forecasts from numerical weather prediction (NWP) data (that is, numerical predictions of wind speed, precipitation, temperature, and other meteorological parameters). SumTime-Mousam is currently being operationally used by Weathernews (UK) Ltd, whose forecasters manually post-edit the computer generated forecasts before releasing them to the ultimate users.

The analysis presented here is mostly based on two corpora created for SumTime-Mousam:

– A corpus of 1045 human-written forecasts, together with corresponding NWP data [15]. An extract from this corpus is shown in Figure 1; the NWP data corresponding to the WIND (10M) statement of Figure 1 is shows in Table 1.

From SumTime-Mousam
FORECAST 00-24 UTC Fri 4-Jul 2003
WIND(10M): NW 15-20 INCREASING 21-26 BY MIDDAY THEN EASING 18-23 BY MIDNIGHT.
SIG WAVE: 1.5-2.0 MAINLY NW SWELL RISING 2.5-3.0 BY AFTERNOON THEN FALLING 2.0-2.5 BY MIDNIGHT.

Post-edited
FORECAST 00-24 UTC Fri 4-Jul 2003
WIND(10M): NW 15-20 INCREASING 21-26 BY MIDDAY THEN **DECREASING** 18-23 **IN THE EVENING**.
SIG WAVE: 1.5-2.0 RISING 2.5-3.0 THEN FALLING 2.0-2.5.

Fig. 2. Extract from post-edit corpus, for forecast issued on 30-Jun-03 (near-synonym changes in **BOLD**)

– A corpus of 2728 post-edited forecasts: this includes the forecast produced by SumTime-Mousam from this data, the forecast actually sent to clients after human post-editing, and the source NWP data. An example of a computer-generated and post-edited forecast is shown in Figure 2. A smaller version of this corpus is described by Sripada *et al* [16].

The post-edit example of Figure 2 illustrates some of the near synonym choices that SumTime-Mousam must make. In this case, the forecaster has replaced *easing* by its near-synonym *decreasing*, and has also replaced *by midnight* by its near-synonym *in the evening*.

3 Analysis of Human Written Forecasts

We analysed near-synonym choice in wind statements in our corpus of manually written forecasts. To do this, we manually grouped words into near-synonym clusters, each of which had a core meaning (similar to Edmonds and Hirst [4]). If a word's meaning was context dependent or otherwise varied, we attempted to disambiguate each usage of the word (for example, time phrases were disambiguated as described in Reiter and Sripada [12]); clusters only contained instances instances where the word was used with the cluster's meaning. Examples of clusters include

– *easing, decreasing, falling*: verbs indicating wind speed is decreasing
– *by midnight, by late evening, later-0000* (that is, *later* used to mean 0000), *by-end-of-period-0000, by-evening-0000, in-the-evening-0000*: time phrases meaning 0000.
– *then, before, and-SEQ*: connectives used to mean temporal sequence.

There were 14 clusters in all.

For each cluster, we extracted from our corpus all wind statements which used the words in that cluster. This resulted in an average of 449 instances of each cluster (individual cluster size ranged from 61 instances to 1106 instances). We then used Ripper [2] and C4.5 [8] (as implemented in Weka's [19] J4.8 classifier) to learn classification rules that predicted which near-synonym in a cluster was used in each text. We experimented with different feature sets to determine which information was most useful in predicting near-synonym choice:

- *semantic* features: change in wind direction, change in wind speed, amount of time over which this change took place, and several derived features (such as the change in wind direction divided by the change in wind speed).
- *author* feature: which forecaster wrote the text.
- *collocation* features: the preceding and subsequent words in the text (numbers and directions were replaced by generic NUMBER and DIRECTION symbols).
- *repetition* feature: the previous word of this type (e.g., verb) in the text.
- *surface* features: the number of words in the sentence, the number of words in the phrase that contained the near-synonym, and the position of the word and phrase in the sentence.
- *temporal* features: the day and hour the forecast was issued on, and how far in the future the prediction was from the forecast issue date.

The above features were chosen on the basis of knowledge acquisition activities with forecasters and small-scale pilot experiments. For the purposes of this paper, all feature sets except 'semantic' are considered to be contextual.

Ripper and C4.5 gave very similar results, below we report only the ones for C4.5. All error rates are computed with 10-fold cross-validation.

3.1 Verbs

Our analysis of the impact of different feature sets in predicting verb near-synonym choice is shown in Figure 2. We show the error rates of classifiers built with no features (baseline), each individual feature set, each combination of author (best individual feature set) with another feature set, and all features. In this analysis we have ignored conjoined verbs, such as *backing and easing*.

Author is clearly the most powerful feature set; it halves classification error, from 16% to 8%. For example, some forecasters preferred to use *rising* as a wind-speed-increase verb, and others preferred to use *increasing*; this simply reflects personal idiosyncracies and writing styles.

Semantic features by themselves were of no help, but when added to the author feature the error rate went down a bit, from 8% to 7%. For example, again looking at wind-speed-increase verbs, one forecaster preferred *freshening* if the wind speed was still moderate (20 knots or less) even after it increased, and *increasing* otherwise. No other forecaster did this. In other words, while individual authors sometimes associated fine-grained semantic connotations with words, these seemed to be idiosyncratic and not shared by the group as a whole.

Table 2. Verb classifier error rates, by feature sets used

features used	error
none (baseline)	16%
author	8%
collocation	14%
repetition	16%
semantic	16%
surface	16%
temporal	15%
author, collocation	8%
author, repetition	8%
author, semantic	7%
author, surface	8%
author, temporal	8%
all	6%

3.2 Connectives

We analysed connectives in a similar fashion to verbs, but due to space limitations we cannot present detailed results here. The baseline (no feature) connective classifier had a 22% error rate. The most useful single feature set was collocation; a classifier based on collocation had a 16% error rate (for example, *and* is strongly preferred for a sequence connective if the subsequent word is *later*). Adding the repetition feature set improved error rate to 14%; essentially forecasters like to vary the connectives used in a sentence.

Adding further feature sets did not significantly improve classification performance. We did note some cases of individual preferences; for example, when authors needed an alternative to *then* to avoid repetition, most used *and* but one used *before*. However, such cases were not common enough to significantly influence overall classification accuracy.

We also attempted to learn a classifier that predicted the punctuation associated with a connective. The baseline classifier (just connective) had a 30% error rate; a classifier based on connective and author had a 5% error rate; and a classifier based on connective, author and surface features had a 4% error rate.

3.3 Time Phrases

The baseline (no feature) classifier for time phrases had a 67% error rate. The most useful single feature for classification was again author; this reduced error rate to 52%. Adding information about the position of a phrase in a sentence further reduced error rate to 48%; for example, one author used *by midnight* to refer to 0000 in the middle of a sentence, but *later* to refer to this time at the end of a sentence.

Adding semantic information did not further improve the error rate. We did notice a few cases where semantics seemed to play a role; for instance one forecaster seemed to prefer *by afternoon* for 1500 when the wind was changing slowly, but *by mid afternoon* when it was changing rapidly. But as with the impact of author on connective choice (see above), these effects were small and had no significant impact on overall error statistics.

The classifier error rate was 46% with all the features sets included.

3.4 Discussion

Our analysis suggests that the author feature is overall the most powerful predictive feature in our set. In other words, the idiosyncracies and preferences ('idiolect') of individual authors has a strong influence on near-synonym choice. Semantics plays little role except in verb choice, but even here its effect is author-dependent. Other contextual features also play a role, including collocation, lexical repetition, and the position of the phrase in the sentence.

Of course, our classifiers still have a high error rate, and it is possible that this is due to a semantic feature which we have omitted from our feature set; in other words, perhaps semantics would be more significant if we used a different feature set. We also of course are working in a single domain (and in a sublanguage), and perhaps different results would be obtained in other domains.

4 Post-edit Analysis

Our post-edit corpus shows how forecasters have edited computer generated texts. These texts were generated using a version of SumTime-Mousam that always chooses the same member of a near-synonym cluster, usually the one that was most common in our corpus analysis. We are currently analysing cases where forecasters have replaced a word by one of its near-synonyms. This work is not complete, but we present some initial results for verb choice below. As above, all figures cited below are for wind statements.

4.1 Verbs

The only case where forecasters regularly post-edited a verb into another verb in the same cluster was changing *easing* to *decreasing*. This happened in 15% of cases. Individual differences are very striking. Of the 9 forecasters for which we have post-edit data, 5 changed *easing* to *decreasing* less than 5% of the time, 2 made this change over 90% of the time, with the remaining two in between (30% and 75%). A classifier (which predicts when *easing* is changed to *decreasing*) built on the author feature has a 5.5% error rate.

We were especially surprised by forecaster F5, who changed *easing* to *decreasing* in 92% of cases. We have data from him in our manual corpus, and in this corpus he used *easing* 69% of the time, and *decreasing* only 30% of the time. However, as noted in Reiter and Sripada [12, Figure 2], F5's behaviour in

this respect may have changed over time. The manual corpus was collected from July 2000 to May 2002, and while at the beginning of this period F5 preferred *easing*, at the end of this period he preferred *decreasing*. Since the post-edit corpus was collected in 2003, this behaviour change may explain the above discrepancy. In other words, not only do individuals have idiosyncratic preferences about near-synonym choice, but these preferences may change over time.

We also asked the forecasters (anonymously) about the *easing* vs *decreasing* choice. The comments received included

1. "Personally I prefer *decreasing* to *easing*"
2. "I tend to think of *easing* being associated with a slower decrease and or perhaps with lower wind speeds or heights"
3. "*Easing* is used when trying to indicate a slight decrease when condition are bad ... it is not used when conditions are quiet"
4. (from forecast manager) "On the whole it seems safer to say *decreasing*"

Note that (2) and (3), which are exactly the sort of subtle semantic differences we expected to find between near-synonyms, are in fact contradictory. The forecaster who said (3) associated *easing* with bad weather, which generally means high wind speeds; while the forecaster who said (2) associated *easing* with low wind speeds. This supports the evidence from Section 3 that subtle semantic differences can be idiosyncratic.

Comment (4), that *decreasing* is the safest choice, presumably because it has the fewest connotations, is interesting. This is supported by another puzzling fact, which is that the *increasing* was edited into a near-synonym (*rising* or *freshening*) in only 1% of cases. Yet in the manually written forecasts, *increasing* was less dominant in its cluster than *easing*; *increasing* was used in 58% of cases for wind-speed-increase, whereas *easing* was used in 71% of cases for wind-speed-decrease. One explanation is that *increasing* (unlike *easing*) is 'safe' in the sense of comment (4), and hence there is no need to change it. Safety is perhaps another factor that should be considered in near-synonym choice.

4.2 Other Types of Words

We have not yet analysed our post-edit data for the other types of near-synonym choices. However, when we asked forecasters in general about problems with SumTime-Mousam's output texts, the only other comment relevant to near-synonym choice was variation in connectives (Section 3.2).

We have not noticed such variation in any other type of word, in either the manual corpus or the post-edit corpus. So variation (at least in this domain) seems important in connectives, but not other types of words.

5 Future Work

Our work to date has focused on understanding how writers choose between near-synonyms. We hope in the future to investigate how the near-synonym

choice affects readers. For example, it may make sense to prefer high-frequency words, because such words are usually read faster [7]. This may be especially important in texts intended for poor readers [3, 18]. It may also make sense to prefer words which mean the same thing to all readers; for example to express 0000 as *by midnight* (which means 0000 to everyone) instead of *by evening* (which means 0000 to some people and 1800 to others) [12]. This is related to the idea of choosing 'safe' words mentioned in Section 4.1.

We also plan to empirically analyse lexical choice in other domains, in particular textual descriptions of medical data (using the corpus from Alberdi *et al* [1]), in a similar manner. We would also like to empirically investigate other types of microplanning choices, such as aggregation.

6 Conclusion

When we started this work, we expected to find that near-synonym choice was mostly determined by semantic details and connotations, along the lines of the comments about *easing* and *decreasing* made in comments (2) and (3) of Section 4.1. In fact, however, our empirical work suggests that near-synonym choice is mostly influenced by contextual information, especially author. Furthermore, when semantics does play a role in near-synonym choice, it often does so in a highly author-dependent way; in other words, semantic connotations often seem to be idiosyncratic and not shared by a linguistic community. Of course we have only looked at one domain, perhaps other domains are different.

From a practical NLG system-building perspective, our current thinking is that in general it probably is not worth trying to choose between near-synonyms on the basis of semantic differences. Instead, the system-builder's priority should be a good understanding of the impact of contextual factors such as collocation, repetition, and individual preferences on near-synonym choice; he or she may also wish to consider safety (chance of misinterpretation). At least in the short term, we believe that a better understanding of these factors may be the best way to improve near-synonym choice in NLG systems.

Acknowledgements

Our thanks to the many individuals who have discussed this work with us, of which there are too many to list here. Special thanks to the forecasters and meteorologists at Weathernews, without whom this work would have been impossible! This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC), under grant GR/M76881.

References

1. Alberdi, E., Becher, J., Gilhooly, K., Hunter, J., Logie, R., Lyon, A., McIntosh, N., Reiss, J.: Expertise and the interpretation of computerized physiological data:

implications for the design of computerized monitoring in neonatal intensive care. International Journal of Human-Computer Studies **55** (2001) 191–216
2. Cohen, W.: Fast effective rule induction. In: Proc. 12th International Conference on Machine Learning, Morgan Kaufmann (1995) 115–123
3. Devlin, S., Tait, J.: The use of a psycholinguistic database in the simplification of text for aphasic readers. In Nerbonne, J., ed.: Linguistic Databases. CSLI (1998)
4. Edmonds, P., Hirst, G.: Near-synonymy and lexical choice. Computational Linguistics (2002) 105–144
5. Fawcett, R., Tucker, G., Lin, Y.: How a systemic functional grammar works: the role of realization in realization. In Horacek, H., Zock, M., eds.: New Concepts in Natural Language Generation. Pinter (1993) 114–186
6. Gove, P., ed.: Webster's New Dictionary of Synonyms. Merriam-Webster (1984)
7. Harley, T.: The Psychology of Language. second edn. Psychology Press (2001)
8. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann (1992)
9. Reiter, E.: A new model of lexical choice for nouns. Computational Intelligence **7** (1991) 240–251
10. Reiter, E., Dale, R.: Building Natural Language Generation Systems. Cambridge University Press (2000)
11. Reiter, E., Sripada, S.: Human variation and lexical choice. Computational Linguistics **28** (2002) 545–553
12. Reiter, E., Sripada, S.: Learning the meaning and usage of time phrases from a parallel text-data corpus. In: Proceedings of the HLT-NAACL 2003 Workshop on Learning Word Meaning from Non-Linguistic Data. (2003) 78–85
13. Robin, J., McKeown, K.: Empirically designing and evaluating a new revision-based model for summary generation. Artificial Intelligence **85** (1996) 135–179
14. Sripada, S., Reiter, E., Davy, I.: SumTime-Mousam: Configurable marine weather forecast generator. Expert Update **6** (2003) 4–10
15. Sripada, S., Reiter, E., Hunter, J., Yu, J.: Exploiting a parallel text-data corpus. In: Proceedings of Corpus Linguistics 2003. (2003) 734–743
16. Sripada, S., Reiter, E., Hunter, J., Yu, J.: Generating English summaries of time series data using the Gricean maxims. In: Proceedings of KDD-2003. (2003) 187–196
17. Stone, M., Doran, C., Webber, B., Bleam, T., Palmer, M.: Microplanning with communicative intentions: The SPUD system. Computational Intelligence **19** (2003) 311–381
18. Williams, S., Reiter, E., Osman, L.: Experiments with discourse-level choices and readability. In: Proceedings of the 2003 European Workshop on Natural Language Generation. (2003) 127–134
19. Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann (2000)